分　类　号：TP399　　单位代码：10183
研究生学号：2013544087　　密　　级：公　开

吉 林 大 学

硕士学位论文

(专业学位)

基于胃癌耐药的调控网络研究

Research on Regulatory Networks Based on Gastric Cancer's Drug-resistance

作 者 姓 名：魏立艳

类　　别：工程硕士

领域（方向）：软件工程

指 导 教 师：刘元宁　教授

培 养 单 位：软件学院

2016 年 5 月

基于胃癌耐药的调控网络研究　魏立艳　吉林大学

基于胃癌耐药的调控网络研究

Research on Regulatory Networks Based on Gastric Cancer's Drug-resistance

作 者 姓 名：魏立艳

领域（方向）：软件工程

指 导 教 师：刘元宁　教授

类　　别：工程硕士

答 辩 日 期：2016 年 5 月 28 日

吉林大学硕士学位论文原创性声明

本人郑重声明：所呈交的硕士学位论文，是本人在指导教师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：2016 年　月　日

《中国优秀博硕士学位论文全文数据库》投稿声明

研究生院：

本人同意《中国优秀博硕士学位论文全文数据库》出版章程的内容，愿意将本人的学位论文委托研究生院向中国学术期刊（光盘版）电子杂志社的《中国优秀博硕士学位论文全文数据库》投稿，希望《中国优秀博硕士学位论文全文数据库》给予出版，并同意在《中国博硕士学位论文评价数据库》和 CNKI 系列数据库中使用，同意按章程规定享受相关权益。

论文级别：■硕士 □博士

学科专业：软件工程

论文题目：基于胃癌耐药的调控网络研究

作者签名：　　指导教师签名：

2016 年　月　日

作者联系地址（邮编）：吉林大学软件学院(130012)

作者联系电话：13500892674

摘　要

基于胃癌耐药的调控网络研究

胃癌，高居世界肿瘤死亡率第二位，而影响胃癌治疗疗效的关键因素在于多药耐药。胃癌耐药的发生机制复杂，是涉及多个蛋白编码基因、多条信号通路的复杂调控网络共同作用的结果，针对任何耐药相关分子的单一干预都难以有效逆转胃癌耐药。基因表达，作为细胞生命活动的一个最基本过程，其内容可概括为 DNA 转录为 mRNA，然后再翻译为蛋白质，而转录阶段则是基因表达的核心步骤，所以一个基因表达的成功与否大大取决于转录阶段，这也因此突显了 mRNA 的重要性。本梯队合作单位第四军医大学研究发现，在胃癌耐药细胞中存在着大量异常表达的 mRNA，并通过对这些 mRNA 表达水平的测量，得到了大量差异表达基因数据，而这些差异表达的基因数据为胃癌多药耐药研究提供了重要线索。

但胃癌耐药中特异表达的基因数量众多，单纯依靠传统生物实验逐条敲出、抑制的研究方法消耗大量时间、物力，且不具有良好的可操作性。生物信息学则可以通过合理的数学建模、利用强大的计算资源，从海量的信息中挖掘出研究人员所关切的、有研究价值的数据，合理的缩小研究对象范围。因此，本文通过合理的建模并结合网络模型，对合作单位提供的大量胃癌耐药差异表达基因数据进行相关统计分析、筛选、挖掘，以从中找到与胃癌多药耐药相关的有价值数据。首先，本文对胃癌耐药差异表达基因数据进行了处理，主要从数据冗余和标注错误两类问题上对数据中存在的不确定错误或噪声进行了处理以提高数据的可靠性，之后对处理后的数据依据注释内容进行了初次分类，并在初次分类基础上，用聚类算法对各分类数据再进行功能性聚类，从而将每分类数据再分为若干个小规模类；然后，在功能性聚类后数据基础上，通过本文挖掘出的各基因间的相互作用关系建立各分类初级基因调控网络，随后用本文设计的评分机制对各分类初级基因调控网络中的各节点进行打分，并依据分值对节点排序，最后依据排序后的节点选取适当的核心节点，进而以选取的这些核心节点为中心用本文设计的模块挖掘方法进行核心网络模块挖掘，接着，对挖掘出的核心网络模块进行分析。经过相关分析后，本文得出 20 个核心基因以及它们的附属基因所组成的网络模块是与胃癌多药耐药相关的有价值数据。

这种以生物信息学为基础，从网络模型角度出发，基于胃癌耐药的调控网络计算模型研究，为胃癌的多药耐药研究提供了新的思路。同时本文也对促进胃癌耐药机理研究、临床治疗、转录组学研究等科学领域都有着重要的理论与现实意义，对其他癌症耐药研究也具有借鉴作用。

关键词：

胃癌，多药耐药，差异表达基因，调控网络

Abstract

Research on Regulatory Networks Based on Gastric Cancer's Drug-resistance

Gastric cancer, the world's second highest cancer mortality, and the key factor affecting the therapeutic efficacy of gastric cancer is multi-drug resistance. Because the drug-resistance mechanism of the gastric cancer is complex and it is combined action result of complex regulatory networks involving multiple protein-coding genes, and multiple signaling pathways, intervention targeting at anyone resistance molecule is hard to reverse the drug-resistance of gastric cancer. Gene expression, as a fundamental process of cellular life activities, Which can be summarized as DNA transcribed into mRNA, and then translated into protein, and transcription stage is the core step of gene expression, so the success or failure of gene expression greatly depend on the stage of transcription, and therefore it also highlights the importance of the mRNA. Our cooperation unit fourth military medical university found that there were a lot of mRNA abnormal expression in gastric resistant cells, and by measuring the expression level of these mRNA got a large number of differentially expressed genes data, and these differentially expressed genes data provide important clues for the study of multi-drug resistance of gastric cancer.

However, a great number of genes are expressed specific in the drug-resistance of gastric cancer, solely using the traditional research methods that knockout and inhibit the gene one by one consume a lot of time and resources, poor feasibility. Bioinformatics can be through reasonable mathematical modeling, using powerful computing resources excavate data from vast amounts of information which are concerned by researchers and have research value, reasonable narrowing the scope of the study. Therefore, this paper through reasonable modeling and combine with network model, analyze, screen, excavate differentially expressed genes data of gastric cancer's drug-resistance which were provided by our cooperation unit, in order to find valuable data associate with multi-drug resistance of gastric cancer. First, this paper deal with these differentially expressed genes data of gastric cancer's drug-resistance, mainly from data redundancy and annotation errors two types of problems that exist uncertain errors or noise are processed to improve the reliability of data, after these processed data based on annotation contents are first classified, and then on the basis of initial classification, each classification will be do functional clustering by using clustering algorithm, thereby each

classification data can be divided into a number of small size class. Then, on the basis of function clustering data, this paper establish each classification's primary gene regulatory network by interactions which were excavated between genes, then measure the value of each node in each classification's primary gene regulation network by scoring mechanism, and sort these nodes according to scores, finally according to these sorted nodes, we select appropriate core nodes, and then we design core network construction module method for these selected core nodes, next, analysis core network modules which have been excavated. After correlation analysis, this paper get the conclusion that network module composed of 20 core genes and their accessory genes are valuable data associate with multi-drug resistance of gastric cancer.

On the basis of bioinformatics, from the perspective of the network model, calculation model research on regulatory networks based on gastric cancer's drug-resistance, provides a new idea for the research of multi-drug resistance of gastric cancer. At the same time, this paper also has important theoretical and practical significance for promoting gastric cancer's drug-resistance mechanism study, clinical treatment, and transcriptome research field, and even for other types of cancer drug resistance research also has reference.

Keywords:

gastric cancer, multi-drug resistance, differentially expressed genes, regulatory networks

目　录

第 1 章　绪　论

1.1　选题依据

胃癌，高居世界肿瘤死亡率第二位，同时也是我国致死人数较多的癌症之一，其发病率在我国居前三位[1]。而与发达国家相比不同之处在于，我国每年新发胃癌病例大多已处于晚期。对于晚期胃癌的治疗，化疗则为主要的治疗手段，因为化疗不仅可以使肿瘤的体积变小，还可以抑制其生长。然而临床研究数据表明，胃癌患者对于目前临床常用的以细胞毒药物为主的化疗方案的总体反应率仅为 40%。而造成这一现象的一个重要原因是肿瘤细胞的多药耐药（multi-drug resistance, MDR），即胃癌细胞同时对多种结构和作用机制不同的化疗药物产生交叉耐药。针对胃癌的多药耐药，国内外学者开展了深入研究，也发现了一批胃癌耐药相关分子[2-3]，包括 ATP（三磷酸腺苷）依赖的药物转运蛋白，药物代谢酶类，药物作用靶标以及细胞凋亡相关分子等。然而，胃癌多药耐药的发生机制复杂，涉及的分子具有多样性，而针对任何耐药相关分子的单一干预都难以有效逆转胃癌多药耐药。

RNA（核糖核酸）是一个种类繁多，功能广泛的复杂的大分子体系，它是连接 DNA（脱氧核糖核酸）与蛋白质的桥梁，通常控制着基因的表达。正因为基因表达的存在才有了生命，而它也是细胞生命活动的最基本过程。基因的表达过程也极为复杂，其主要内容可概括为：DNA 转录为 mRNA（信使 RNA），然后再翻译成蛋白质，而转录阶段则是基因表达的核心步骤，所以一个基因表达的成功与否大大取决于转录阶段，这也因此突显了 mRNA 的重要性。本梯队合作单位第四军医大学近些年通过生物实验研究亦发现，在胃癌耐药细胞中存在着大量异常表达的 mRNA，并通过对这些 mRNA 表达水平的测量，得到了大量差异基因表达数据，筛选并鉴定这些差异表达基因，探究其在胃癌 MDR 机制中的作用具有重要意义。

但胃癌耐药细胞中特异表达的基因数量众多，若单纯依靠传统生物学实验采取逐条敲出、抑制的研究方法无论从时间还是人力、物力、成本的角度考量，都不具有很好的可操作性[4]。生物信息学则可以通过合理的数学建模、利用强大的计算资源，从海量的信息中挖掘出研究人员所关切的、有研究价值的数据，从而合理的缩小研究对象范围。而基于胃癌耐药的调控网络生物信息学研究模型至今仍尚未见报道。因此，以计算机技

术为工具，从网络模型角度出发，研究基于胃癌耐药的调控网络计算模型，为生物实验做理论指导胃癌 MDR 研究显得尤为重要。同时，也相信本研究对促进胃癌耐药机理研究、临床治疗、转录组学研究等科学领域都有着重要的理论与现实意义，且对其他肿瘤耐药研究也具有借鉴作用。

1.2　研究背景及国内外研究现状

仅次于肺癌的胃癌，居世界肿瘤死亡率第二位，然而该病在我国以及韩国、日本等国家尤为严重，严重威胁着这些国家人民的生命健康[5]。化疗作为主要的胃癌治疗手段之一，效果往往非常有限，导致这种效果有限性的根本原因是胃癌 MDR 的存在。近些年不断发展的基因、蛋白组学研究技术为胃癌 MDR 研究提供了有利工具，也因此发现并报道了许多与胃癌耐药相关的分子。但由于胃癌发病情况分布的差异，使得国外对该病的 MDR 相关研究报道相比国内较少。而国内的胃癌 MDR 研究则主要以生物实验为主，即通过生物实验最终证实一个或几个分子与胃癌 MDR 相关。最具代表性的则是一直致力于胃癌耐药研究的第四军医大学，他们也通过生物实验筛选得到了大量与胃癌耐药相关的分子[6-8]，而且部分分子均已生物实验验证。

目前针对胃癌 MDR 的研究多以生物学研究为主，而生物学因受人力、物力、成本等因素的限制只能单一地研究单个或几个分子，但胃癌 MDR 是涉及多基因，多因素，以及多步骤的一个复杂过程，无疑单纯的依靠生物实验研究对于复杂而庞大的胃癌耐药调控网络的认识是有限的。生物信息学，作为一门涉及生物学、数学、计算机科学的交叉学科，它主要是对生物信息进行获取、加工、存储、分析解释，为胃癌的 MDR 研究提供了新的方法。

到目前为止，生物信息学主要经历了三个发展阶段[9]，而每个阶段又都有其各自的研究内容，为此本文对其进行了总结，详见图 1.1。

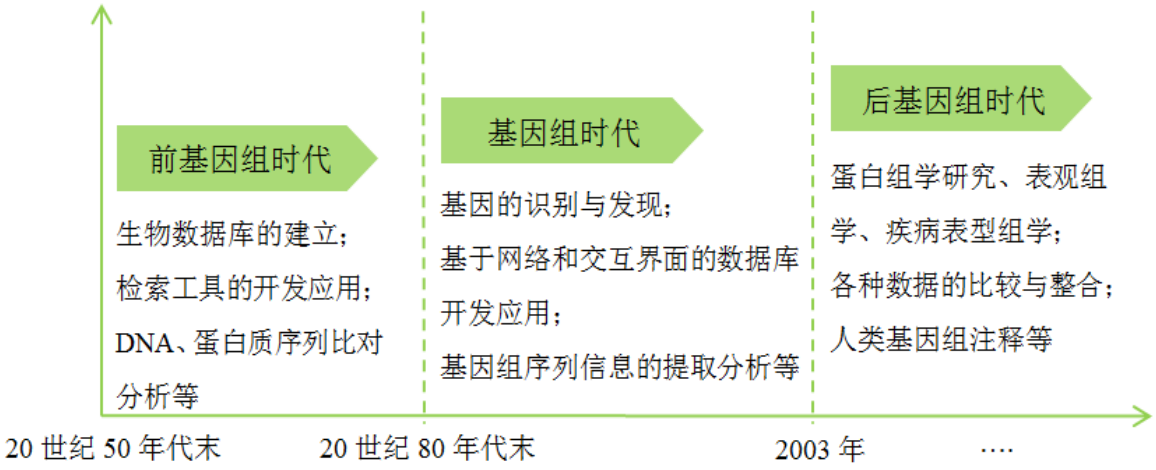


图 1.1　生物信息学三个发展阶段

国外生物信息学研究：1. 研究机构，先后成立了美国 NCBI（国家生物技术信息中心），欧洲 EBI（欧洲生物信息学研究所），日本 CIB（信息生物学中心），并同时成立了 Gene Bank/EMBL/DDBJ 国际核酸序列数据库。2. 数据分析技术，早在 1970 年 Needleman[10]就发表了关于两序列比对算法的文章，也因此，Computer Methods and Programs in Biomedicine 期刊诞生；1975 年 Pipas[11]率先提出使用计算机技术预测 RNA 二级结构；1990 年 NCBI 研发了核酸/氨基酸序列相似性比较程序 blast；2000 年人类基因组计划“阿波罗计划”完成；之后越来越多研究分析基因组间关系的方法不断涌现[12]，如网络模型方法、机器学习方法等。3. 相关专业生物信息学刊物、生物信息学网站及测序技术亦随着生物信息学的不断深入研究而不断进步和完善，而测序技术的不断发展又为蛋白组学研究、转录组研究、基因组研究和功能基因组研究提供了数据支持[9]。

国内生物信息学研究：1. 研究机构上以 IPC（北京大学生物信息服务中心）为代表，且该机构也已建立 EMBL 中国镜像数据库。2. 蛋白组学研究，其研究内容有蛋白质的功能，蛋白质一级结构到四级结构的预测和蛋白质结构异常所引发的疾病等。3. 转录组研究，主要以非编码 RNA 研究为主。4. 基因组研究，包括通过测序和拼接技术对基因注释、重注释，基因的差异表达分析研究[13]，以及基因注释和功能研究促进基因相互作用的研究[14]。5. 功能基因组研究，该技术主要为通过基因芯片及转录组测序等技术，研究基因在特定组织特定时期的表达信息，并依据这些表达信息发展算法进行进一步分析，从而最终获得复杂生物网络[15]。例如：蛋白相互作用网络，基因表达调控网络等。

功能基因组学研究使得研究人员对基因层次的研究更为深化，人类也可以更好地从微观分子层面上探索生物基因的奥秘。基因调控网络作为功能基因组学研究的重要内容亦被广泛研究。经过多年研究，基因调控网络相关研究也取得了突破性成果。

每个基因的表达都受一个或多个基因的协同控制，而该基因又对另外的某个基因或与其他基因协同作用共同控制某一基因的表达，这种相互作用、相互约束的复杂关系便组成了基因调控网络[16]，它从系统的观点出发，通过对网络中的各分子及其间的作用关系研究来探索人类生命现象，显然，单一地研究某个基因是不可行的。基因调控网络是利用生物信息学方法及技术研究数据间的复杂网络关系，而基因间的相互作用关系则是建立网络的纽带。

基因调控网络要想从生物信息学角度研究需要一个前提假定，即任意两个基因的序列谱相似，则它们相互作用，且功能上也相似[17]；基因的表达模式相同则表达过程可能也相同[18]。该假定主要是因为我们在建立网络时所使用的数据只是一类表达浓度数据，显然通过这类数据是无法得出各基因间的相互作用关系的，因此，需要这样一个假定让我们推出各基因间的相互作用关系，从而建立基因调控网络。

基因调控网络通常有下述四个特性：1. 复杂性，每个基因大多都受一个或多个基因的影响。相关研究也表明，任一基因均与 4-8 个基因之间存在着联系[19-20]，有的甚至能与 10 个以上基因之间存在相互作用关系[21-22]。2. 稳定性，基因的表达必须有一个相对稳定的网络系统环境。3. 动态性，各基因组成的网络系统不是一成不变的而是经常变化的，一旦网络系统发生变化，网络中的各基因也会在不同时期、特定细胞和特定细胞环境下受到影响，进而作用到基因调控网络的各个节点上，同时影响基因的相应表达。4. 功能模块性，功能相同或功能类似的基因往往会聚拢在一起形成一个模块，而这些聚拢在一起的基因要么是按照一定的顺序表达，要么是同时表达。

基因调控网络模型研究始于 20 世纪 60 年代，随着研究的不断深入和相关技术的不断发展，网络模型构建研究及算法的探索也得到了极大地提升。但因国内相关研究起步较晚，所以，现多以国外研究为主。而到目前为止也已出现了大量相关网络构建模型和算法，如微分方程模型、布尔网络模型、线性组合模型、贝叶斯网络模型等。

微分方程模型是众多模型中应用最为广泛的一种，它以 RNA、蛋白质及其他相关元素的浓度为研究对象，该模型的各变量的取值是与时间相关的。该模型较早有 Chen 等人[23]应用，他们使用线性常微分方程来描述整个生物系统，随后 Tominaga[24]等人提出了一种基于遗传算法的微分方程系统的 S-system 模型，该模型中所引入的遗传算法可以优化系统的大量参数。由 Kauffman[25]引入的布尔网络模型，是最为简单的一种调控网络模型，在模型中每个基因只有 ON，OFF 两种状态，即开和关，基因间的相互作用关系亦遵循布尔规则，随着该模型研究的不断深入，布尔网络得到了进一步扩展，出现了时

序布尔网络模型，该模型可处理多于一时间单位跨度的基因间的依赖性，随后 Shmulevich 等人[26-27]又将马尔可夫链与布尔网络相结合用以解决概率框架下的不确定性问题，他们同时还引进了概率布尔网络模型。线性组合模型，一种连续的调控网络模型，该模型假设基因间的相互作用是线性非瞬时的。针对此模型先前的研究方法多为先把众多基因全局聚类，然后再寻找类间的调控网络，而类内基因间的作用关系则用其他知识辨认。随后张晗等人[28]更好地发展了此方法，他们先用线性模型建立基因类间调控网络，然后再用多元回归模型建立基因间调控网络，从而获得了更有意义的结果。贝叶斯网络模型[29]，一种重要的概率模型，有条件概率分布和网络结构两部分组成，其中，贝叶斯公式是概率网络的基础。该模型一般适用于分析反映复杂生物系统内部关系的基因表达数据，通常分为静态贝叶斯网络模型和动态贝叶斯网络模型两类。

1.3　本文的研究目的与意义

作为在我国呈现诊断率低，死亡率高特点的胃癌，MDR 的产生是制约胃癌化疗疗效的主要问题，而癌症时基因的表达水平与非癌时肯定是不同的，因此，差异表达基因的筛选和鉴定为胃癌耐药研究提供了重要线索。但目前针对胃癌 MDR 的研究主要以生物学实验为主，因受成本等各种因素的限制使得研究的基因数量往往有限，而这对于涉及多基因，多因素以及多步骤的胃癌 MDR 研究显然远远不够。

为此，本文借助生物信息学方法，通过合理的建模并结合网络模型，对合作单位提供的大量胃癌耐药差异表达基因数据进行相关统计分析、筛选、挖掘，以从中找到与胃癌 MDR 相关的有价值数据。首先，本文对胃癌耐药差异表达基因数据进行了处理，主要从数据冗余和标注错误两类问题上对数据中存在的不确定错误或噪声进行了处理以提高数据的可靠性，之后对处理后的数据依据注释内容进行了初次分类，并在初次分类基础上，用聚类算法对各分类再进行功能性聚类，从而将每分类数据再分为若干个小规模类；然后，在功能性聚类后数据基础上，通过本文挖掘出的基因间的相互作用关系建立各分类初级基因调控网络，随后用本文设计的评分机制对各分类初级基因调控网络中的各节点进行打分，并依据分值对节点排序，最后依据排序后的节点选取适当的核心节点，进而以选取的这些核心节点为中心用本文设计的模块挖掘方法进行核心网络模块挖掘，接着，对挖掘出的核心网络模块进行分析，从而找到与胃癌 MDR 相关的有价值数据。

这种借助生物信息学方法，从网络模型角度出发，基于胃癌耐药的调控网络计算模型研究，既为胃癌 MDR 生物实验提供了理论指导，也为寻求胃癌耐药机理提供了新的思维方式。同时，本文也对促进胃癌耐药机理研究、临床治疗、转录组学研究等科学领域都有着重要的理论与现实意义，对其他癌症耐药研究也具有借鉴作用。

1.4　本文的组织结构

第1章：绪论。本章主要从选题依据，研究背景及国内外研究现状，研究目的与意义，以及文章的组织结构四方面初步介绍了本文的研究工作。

第2章：胃癌耐药调控网络研究相关知识介绍。本章首先介绍了肿瘤多药耐药基本知识，接着从基因的表达过程，基因的调控机制两方面对调控网络原理进行了介绍，然后又介绍了几种常见的基因调控网络模型，为本文的研究工作奠定了理论基础。

第3章：胃癌耐药差异表达基因数据预处理。本章主要从数据形式及其处理，数据的注释分类，以及功能性聚类三方面对数据集进行了预处理，介绍了如何对数据集中存在的不确定错误或噪声进行处理，如何将数据依据注释内容进行初次分类，以及如何在初次分类基础上使用功能性聚类方法将每类数据再分为若干个小规模类。

第4章：基于胃癌耐药数据的调控网络研究。本章首先对基于胃癌耐药数据调控网络研究的总体流程进行了介绍，然后对调控网络研究的具体设计做了详细介绍，包括如何对各分类基因数据进行融合建立初级调控网络，如何衡量初级网络各节点的重要性从而选取适当的核心节点，如何以核心节点为中心挖掘网络模块，以及对挖掘出的网络模块进行分析，从而最终找到与胃癌 MDR 相关的有价值数据。最后，为了证明本文提出的模型方法的有效性，本文用另一套数据对其进行了测试评估。

第5章：总结与展望。本文的研究为生物实验研究胃癌 MDR 提供了理论指导，也为寻求胃癌耐药机理提供了新的思路，但本文所得出的结果仍需要通过相关生物实验进行验证。同时，本文所提出的研究模型也存在诸多不足，需进一步完善。

第 2 章　胃癌耐药调控网络研究相关知识介绍

2.1　肿瘤多药耐药基本知识

依据药物的响应性可将肿瘤耐药分为天然性耐药和获得性耐药两大类。天然性耐药与治疗所使用的药物无关，它是在尚未使用化疗药物前肿瘤细胞就已具有耐药性；获得性耐药则与治疗所使用的药物相关，它是在化疗过程中肿瘤细胞逐步形成的耐药性。通常，大部分肿瘤细胞的耐药性都属于由化疗药物诱导形成的获得性耐药。更糟糕的是，只要肿瘤细胞对某种药物形成了耐药性，就可以同时形成对其它未曾使用的或结构及作用机制都不相同的多种化疗药物的耐受性，即多药耐药（multi-drug resistance, MDR）性[30]。前期研究也发现，胃癌细胞的获得性 MDR 所占比例达百分之百，而天然性 MDR 只有其一半左右。而通常只要胃癌细胞发生了 MDR，无论采用何种方案治疗，最终的治疗效果都不是很理想。

肿瘤 MDR 的发生机制复杂，主要分为以下两大类：

（1）治疗药物无法有效作用于细胞内靶分子。

a）细胞摄取的治疗药物减少或是细胞对治疗药物的外排增多，其中后者为主要方式；

b）细胞增强了治疗药物的解毒功能；

c）治疗药物靶向了设计之外的分子等。

（2）治疗药物无法与细胞内靶分子有效结合从而杀死或抑制肿瘤细胞。

耐药细胞亚系与其亲本细胞表达谱的差异比较是肿瘤 MDR 相关研究的手段之一。而胃癌 MDR 的相关研究亦通过该方法发现了大量与耐药相关的基因。细胞中不同组分之间不是孤立存在的而是交互作用的，同样，涉及胃癌 MDR 的基因亦不是孤立存在的，它是若干个基因复杂调控网络共同作用的结果。显然，单纯的依靠生物实验研究胃癌耐药调控网络是远远不够的，因此，理论分析并结合计算机方法对调控网络进行研究显得尤为重要。而近些年来，调控网络的相关理论研究也取得了一系列成果，这些成果都将是解决生物学问题的手段。

2.2　调控网络基本原理

2.2.1　基因的表达过程

基因表达，作为一个高度受调控的基本生物过程，它通常指把遗传信息经过转录和翻译转为具有生物活性的蛋白质分子，而这个过程亦是生命体系中最核心、最简约、最本质的规律-生物中心法则。

生物中心法则是所有具有细胞结构的生物必须遵循的法则，它于 1958 年由 Crick 提出，他认为遗传信息的流动方向有两种：第一种，从 DNA 传到 RNA 再传到蛋白质；第二种，从 DNA 传到 DNA 进行自我复制。之后有科学家发现在一种被称为逆转录酶的作用下，RNA 能合成 DNA。同时，另一病毒实验也表明，RNA 具有自我复制功能。因此，1970 年 Crick[31]对生物中心法则做了补充完善，完善后的中心法则见图 2.1。

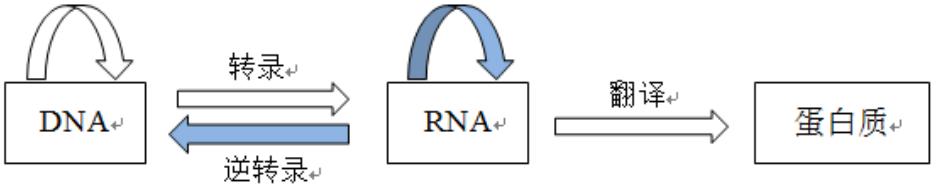


图 2.1　生物中心法则

在完善后的中心法则中，遗传信息的传递过程共包含了五条线路：1. DNA 的自我复制。2. DNA 将遗传信息转录为 mRNA。3. 通过蛋白酶将信息翻译为蛋白质。4. 逆转录酶作用下，RNA 合成 DNA。5. RNA 的自我复制。

中心法则中主要涉及了 DNA、RNA、蛋白质三种大分子，因此，首先需要了解一下这些概念[32-33]。

DNA，脱氧核糖核酸，它是储存遗传信息的最重要的一种具有双链结构的生物大分子，主要存于细胞核中，有 A（腺嘌呤）、G（鸟嘌呤）、C（胞嘧啶）、T（胸腺嘧啶）四种碱基组成。遗传信息往往被编码为核苷酸序列，而核苷酸序列的顺序又决定了细胞内 RNA 及蛋白质的基本结构。

RNA，核糖核酸，一种长链状分子，一般存于细胞质中，有 A（腺嘌呤）、G（鸟嘌呤）、C（胞嘧啶）、U（尿嘧啶）四种碱基组成。通常 mRNA（信使 RNA）、tRNA（转运 RNA）、rRNA（核糖体 RNA）三种主要 RNA 参与蛋白质的合成。

蛋白质，一种有机大分子，它是生命的物质基础，是组成细胞的主要有机物之一。

氨基酸是组成蛋白质的基本单位，而氨基酸又可以分为 20 多种，它们按不同的比例组合并在体内不断代谢与更新。蛋白质共具有一级、二级、三级、四级四个结构，而不同的结构又决定了蛋白质的不同功能。

基因，作为惟一能自主复制、永久存在的单位，它是大分子 DNA 的一段序列，由编码区、调控区两部分组成，通常控制着蛋白的合成，是转录功能的单位，其产物可以是蛋白质，也可以是 RNA，而储藏在任一基因中的生物信息都必须先被转录为 RNA，才能得以表达，这便是基因表达的第一步：转录。

基因转录，它指以 DNA 为模板，RNA 聚合酶为催化剂，按照碱基互补配对原则，合成 mRNA 的过程。它是基因表达的核心步骤，因此，基因的表达成功与否大大取决于该阶段。然而，在遗传信息的传递过程中，基因组遗传信息只是将一部分表达为 RNA，不表达为 RNA 的那部分则为基因表达的调控区。mRNA 主要是把 DNA 某区间段的遗传信息转录下来，然后在细胞质中的核糖体上以其自身为模板控制蛋白的合成。此外，转录过程中还会产生另外一类主要的 RNA：tRNA，它的主要功能是在蛋白质合成过程中运载氨基酸，每一种氨基酸都有相应的 tRNA。

基因翻译，基因表达的第二步，它是指遗传信息以 mRNA 为模板，合成具有特定氨基酸顺序的蛋白质的过程。该过程中参与的主要分子有 mRNA、tRNA、rRNA、蛋白质。基因在翻译水平的调控方式则因所在细胞不同而不同，它大体上主要有以下两方面：1. 依据细胞状态决定是否执行翻译过程，以保证合成大量蛋白质。2. 微调 mRNA 以控制所合成蛋白质的质和量。

蛋白质合成的各阶段都有调控的参与，而每个基因的表达亦有其严格的调控机制，以使其产物达到细胞所需最佳量。通常基因表达的调控是多层次的，一般为上一层基因产物调控下一层基因表达[34]。我们可以把这种调控过程看成由众多基因组成的逻辑网络，而这种网络可简单看作由基因和基因间的连线组成。

2.2.2　基因的调控机制

依据产物的功能可将基因分为结构基因、调控基因两种。结构基因，一类编码蛋白质（或酶）或 RNA 的基因，其功能是把携带的遗传信息转录给 mRNA，然后再以其自身为模板合成具有特定氨基酸序列的蛋白质或 RNA。调控基因，一类调控蛋白质合成的基因，通常对结构基因具有调节作用。它的功能主要是产生一类抑制物，以制约其他

基因的活动，它犹如一个自动控制系统，使得结构基因在需要某种酶时就合成，不需要时就中止。

生物体广泛存在诱导作用、阻遏作用两类调控机制[35]。每种调节类型的诱发剂一般为小分子物质，或酶底物，或酶催化合成的产物。诱导作用是指某种酶或一组酶共同作用促使细胞中某物质合成的能力，而触发此调节类型的小分子物为诱导物。阻遏作用则与诱导作用相反，它是关闭细胞中某物质的合成，触发该调节类型的小分子物为阻遏物，而阻遏物只有在细胞内的特定代谢物质浓度达到一定量时才会被活化。

调控因子，作为基因受调控的主要作用物，一般分为激活蛋白、阻遏蛋白两种。激活蛋白是启动或促进结构基因转录的一种调控蛋白。而阻遏蛋白与激活蛋白相反，它遏止或减弱结构基因的转录。调控因子实现对其他基因表达调控的方式通常有两种：第一种，提高基因的转录水平；第二种，提高 mRNA 的翻译水平。其中，第一种为主要方式。

调控因子对靶基因的调控往往会导致基因转录水平的下降或上升。如果调控因子同缺少其相比使靶基因的表达水平下降了，那么这种调控方式则为负调控，反之为正调控。负调控作用的调控因子为阻遏蛋白，该蛋白主要是抑制转录的开始而不是辅助其开始，因此，当该蛋白失活时，基因反而能正常表达，而当该蛋白存在时，基因却无法正常表达。正调控作用的调控因子为激活蛋白，它与负调控相反，若该蛋白失活，则基因不能正常表达，而当该蛋白存在时，基因正常表达。

调控网络中，每个基因的表达都受一个或多个基因的协同控制，而作为网络中的基本元素，该基因对另外的某个基因或与其他基因协同作用共同控制某一基因的表达。因此，深入研究基因调控网络模型具有重要的理论现实意义。

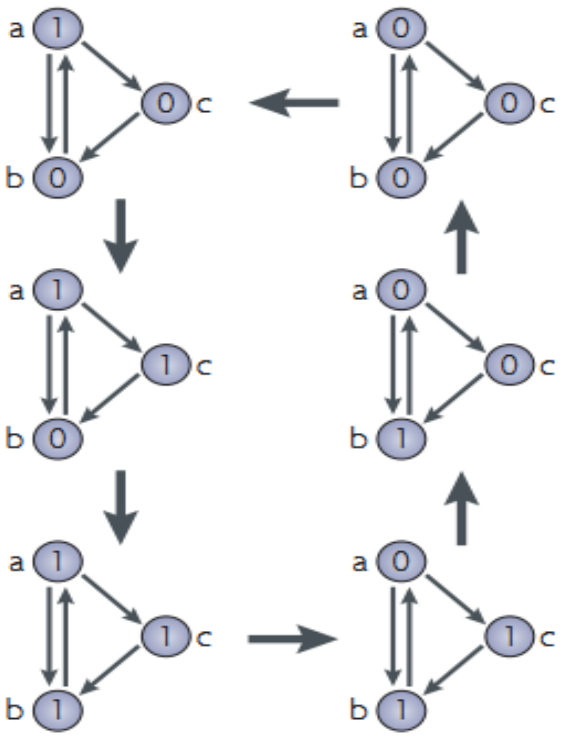
2.3　基因调控网络模型

基因调控网络是由 DNA、RNA、蛋白质及代谢中间物等参与基因调控作用的生物分子与其间的作用关系所形成的网络[34]，它有助于我们更好的从整体水平上理解基因间的相互作用关系，进而探索人类复杂生命现象。到目前为止，已出现多种网络模型，这些模型大致可分为以下三类[36-37]：逻辑模型，定性描述调控网络；持续性模型，主要关注生理过程如何受时间状态转变的影响；单分子水平模型，描述基因调控与时间状态变化间的关系。

2.3.1　逻辑模型

逻辑模型，一种最基本、最简单的建模方法，它使得用户基本了解不同条件下给定网络的不同功能，尽管该模型只能解答定性问题，但正是该定性特性使得此模型灵活，且易于适应生物现象。逻辑模型的典型代表为布尔网络[38-39]，在布尔网络中，每个实体节点有激活（on，1）或非激活（off，0）两种可选状态。每个实体节点状态的更新都由布尔函数依据其他实体节点状态决定。图中所有节点的 0，1 矢量在任一时刻的所有局部状态（实体的状态值）组合构成了系统状态，即全局状态；假定变化是同步的，那么在任一时序，任一实体的状态可根据前一时序它的调控者的状态或调控函数（决定实体状态的规则）决定。此类网络一般用于阐明网络稳态与调控功能间的关系。

以图 2.2 为例，该网络有 a、 b、 c 三个实体，每个实体有 0、1 两种状态，状态之间的转换则依据右边的调控函数。比如，如果 a 状态为 1，c 状态为 0，那么下一时序 b 的状态为 0，细箭头表示每个节点的调控者，粗箭头表示时序步。每三个实体状态组成一个全局状态。该系统周期通过这六个全局状态表示。



Regulation functions

F_b (a,c)

a	c	b
0	0	0
0	1	1
1	0	0
1	1	1

F_a (b)

b	a
0	1
1	0

F_c (a)

a	c
0	0
1	1

图 2.2　布尔网络示例

由于实验证据的不足，以及对系统的不完整理解，使得布尔网络有一定的局限性，为此，Shmulevich 等人对该网络模型进行了扩充，提出了与实际网络更贴近的概率布尔网络。在该网络中每个实体可拥有多个调控函数，每个调控函数依据之前数据的兼容性

会得到一个概率值，在每一个时序步，每个实体会依据所定义的概率随机选择调控函数，从而决定下一时序步状态。

2.3.2　持续性模型

生物实验产生的数值往往是真实连续的而不是离散的。然而，逻辑模型将真实数据离散化，大大降低了数据的准确性。持续性模型[40]则弥补了该不足，它在一个连续的时序表上使用真实的生物数据，形成全局状态和实验数据间的一个直接比较，原理上较逻辑模型更为准确。其典型代表为调控型通量平衡分析模型[41]，该模型旨在集成调控网和代谢网，从而更加综合性的体现生物过程。该模型的最大特点是通过假定其网络系统状态稳定从而认定网络中实体产生的底物浓度也不发生变化，而该网络模型的目标则为结合多条通路流量，寻求最优解。

以图 2.3 为例，该模型包含三个调控基因 r1，r2，r3（方框），代谢过程用圆圈表示，连接代谢过程间的流量为代谢流，即 v1-v8 为代谢流，其中，箭头的指向代表两个因子间的作用方向，模型的目标为求解 v7+v8 的最优解。代谢流 v7 作用于基因 r1，若 v7 的流量不为 0，则 r1 处于激活状态，反之，r1 为非激活状态。r2 和 r3 作用于代谢流 v5，仅当 r2 处于非激活状态，r3 处于激活状态时，v5 的流量为 0，否则 v5 非约束。而当 v5 非约束时，v7+v8 获得最优解，且所有代谢流（除 v6 外）均获得约束值 0.2，其他值则为 0。当 v5 约束值为 0 时，v7 也必须为 0。化学计量矩阵则用以描述代谢过程每个代谢流的消耗过程，其中列表示代谢流，行为代谢过程。而图 2.3 包含 5 个代谢过程，8 个代谢流，所以其化学计量矩阵为 5*8 矩阵，5*8 矩阵中第三列的数值表示第三个代谢流为代谢过程 2 产生一分子则需消耗代谢过程 1 一分子。

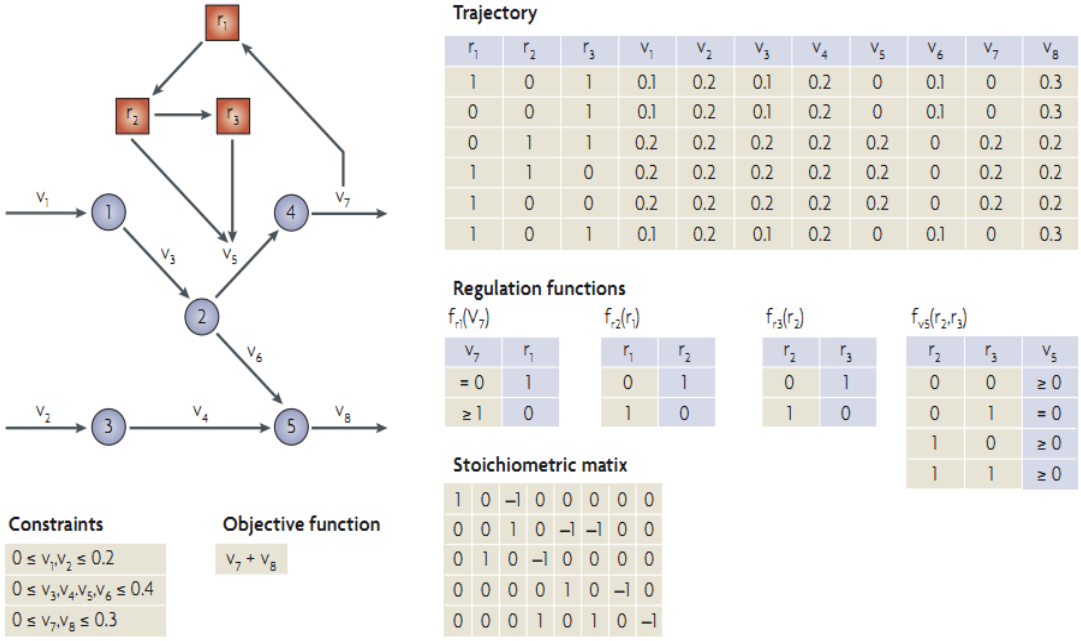


图 2.3　调控型通量平衡分析模型示例

除此之外，持续性模型还包括一类较为常用的模型，常微分方程模型[37]，如图 2.4 所示。模型中的方程描述了每个实体瞬时变化，常微分方程方法也为网络动力学研究提供了更为详尽的信息，然而动力学参数需要高质量的数据，也因此使得此模型目前仅限于部分网络系统使用。

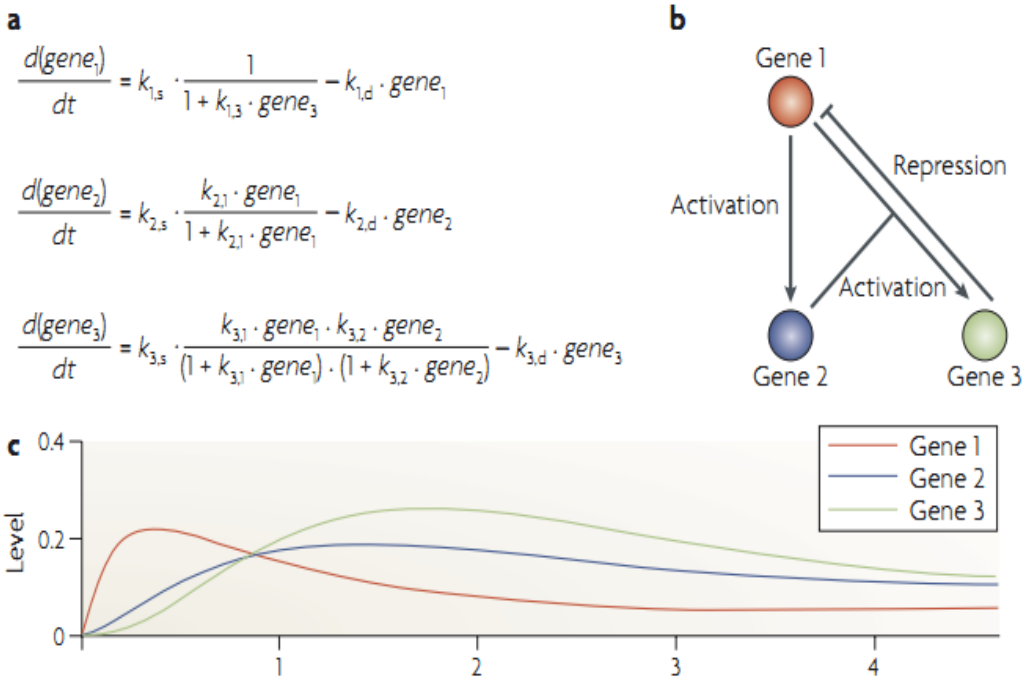


图 2.4　常微分方程模型示例

图中 a 为一个使用常微分方程建模的含三个基因的网络，k 为反应速率常数；b 为

调控关系描述的图形化；c 则为模型的抛物线表示。其中，每个方程体现了基因水平合成、降解的差异变化。gene 1 的表达受 gene 3 抑制，因此，它的表达水平在 gene 3 为 0 时，可能达到一个最高的增长率，该种情况下 $k_{1,s}$（代表合成）需要乘 1。当 gene 3 的水平不再为 0 时，gene 1 的增长水平会慢于 $k_{1,s}$，gene 2 的转录由 gene 1 激活，同样地，gene 3 的转录在 gene 1 和 gene 2 表达水平都非 0 的情况下激活，其关系见 a 面板公式 3。$k_{i,d}$ 为降解速率常数（i 可为 1，2，3），这些公式假定每个转录瞬时翻译，因此，合成常数 $k_{i,s}$ 与转录、翻译均相关。由 c 面板可以看出系统在 4.5 时间单元达到了稳态。

2.3.3　单分子水平模型

每个生物网络都由随机组件构成，因此，相同初始条件下也可能会产生不同行为[42]。当每个种类所涉及的分子数数量巨大时，质量作用定律则可用于精确计算浓度上的变化，且随机效应也会相应减少，而当分子数过小时，则会显著看到随机效应。而通常情况下，调控网络中的分子数量也是较低的，这就使得所建立的调控网络具有随机性。而具有随机模拟特性的单分子水平模型则可以很大程度上提高所建调控网络的准确率。

该模型的代表为 Gillespie 提出的 Gillespie’s 随机模拟算法[43]，该算法需要输入几个物种的初始分子数（如 mRNA 和蛋白质）和反应概率（特定分子在无穷小的时间间隔内组合参与反应的机率）常数，然后通过反应模拟系统网络。该算法主要有以下三个要点：1. 决定下一次反应的发生时间和类型。2. 依据上述反应修改系统状态。3. 继续模拟直到系统处于下一状态。然而，因该算法不仅需要详尽的反应相关理论知识，而且开销费用较高，所以使得此方法的应用效果也较差。

以上不同模型的比较可由图 2.5 给出，从中可以看出，布尔网络是逻辑模型中最纯的一种形式，它们高度抽象，因此需要的数据也最少，也因此使得其只能展示定性动态行为。调控型通量平衡分析产生一个可以与实验测量值相比较的代谢预测，但需要相关生化知识且分析具有挑战性。常微分方程生化机制难以分析。单分子水平模型，一种更为详细的模型，能捕获随机性，但计算开销大。

图 2.5　不同模型图解比较

2.4　本章小结

本章主要介绍了相关调控网络研究知识。包括原发性耐药、获得性耐药的含义，以及肿瘤 MDR 发生的两种机制；接着又从基因的表达过程，基因的调控机制两方面对调控网络原理进行了介绍；然后分别介绍了逻辑模型、持续性模型和单分子水平模型三种调控网络模型。这些相关知识为后续工作提供了理论基础。

第 3 章　胃癌耐药差异表达基因数据预处理

3.1　数据形式及其处理

数据是一切模型和算法的作用对象，没有数据就好比无米之炊，一切都将不可行。构建调控网络所需要的基因数据种类繁多，且形式多样，只要是能用于构建网络模型的生物实验数据都可使用，而基因表达谱则是当前使用最多的一种数据。这种数据类型为完全谱式数据，其通过数字形式提供了某个或某几个基因在某些特定实验生存条件和某些特定表达时间下转录为 RNA 的数量表示，通常为芯片格式。

本文研究所使用的数据集则来自本梯队合作单位第四军医大学，他们以取自国人胃癌标本建系的胃癌细胞株 SGC7901 作为亲本细胞，然后再采用相关生物技术培养了 SGC7901/VCR、SGC7901/ADM 两个稳定的胃癌耐药细胞系，之后他们又利用芯片技术比较了胃癌耐药细胞系与其亲本细胞基因表达谱的差异，从而得到大量差异表达基因数据，而这些数据则是本文研究工作的初始数据集。但是，通常实验方法以及实验环境等客观条件都会或多或少的影响实验结果，从而使得结果集中存在一些不确定的错误或噪声。因此，为了提高数据的可靠性需要对初始数据集进行处理。

经相关统计分析后，发现数据集中主要存在数据冗余和标注错误两类问题，因此本文主要从这两方面对数据进行处理。

（1）数据冗余

对于同一基因，不同的实验结构、权威数据库给出的基因标识符不尽相同。因此，就会出现同一基因但标识符不同的数据冗余现象，为了解决冗余数据问题，本文采用了单联算法 DAVID Gene Concept[44]，该算法是基因组学数据挖掘分析工具 DAVID[45]（注释，可视化和集成发现数据库）的一个子工具。为了改善基因 ID 类型的交叉映射能力，DAVID 采用单联方法从各种公共基因组资源库将成千上万的基因/蛋白标识符凝聚到 DAVID 基因簇，即单联算法 DAVID Gene Concept。

该算法是把冗余 Gene ID 凝聚到 DAVID 基因簇中的一个新型单联算法，DAVID 基因簇作为二次基因簇，通过单联算法合并现有三大重要非冗余基因簇数据库 NCBI Entrez Gene，UniProt UniRef100，PIR-NREF100。对来自上述三种数据库的任何一个基因簇但凡有一个或多个 ID 相同，来自同一物种便认为是同一基因条目，重叠的基因簇

根据单联规则迭代合成一个新的基因簇直到所有最终基因簇或 DAVID 基因稳定为止，并将一个唯一整型数据分配给每一个新形成的 DAVID 基因，同时将其作为集中基因标识符，任何属于同一基因类型的亚型或剪切体都归为同一 DAVID ID，单联算法反复迭代聚合属于相同基因条目中的所有类型基因 ID。图 3.1 为单联算法实现示例。

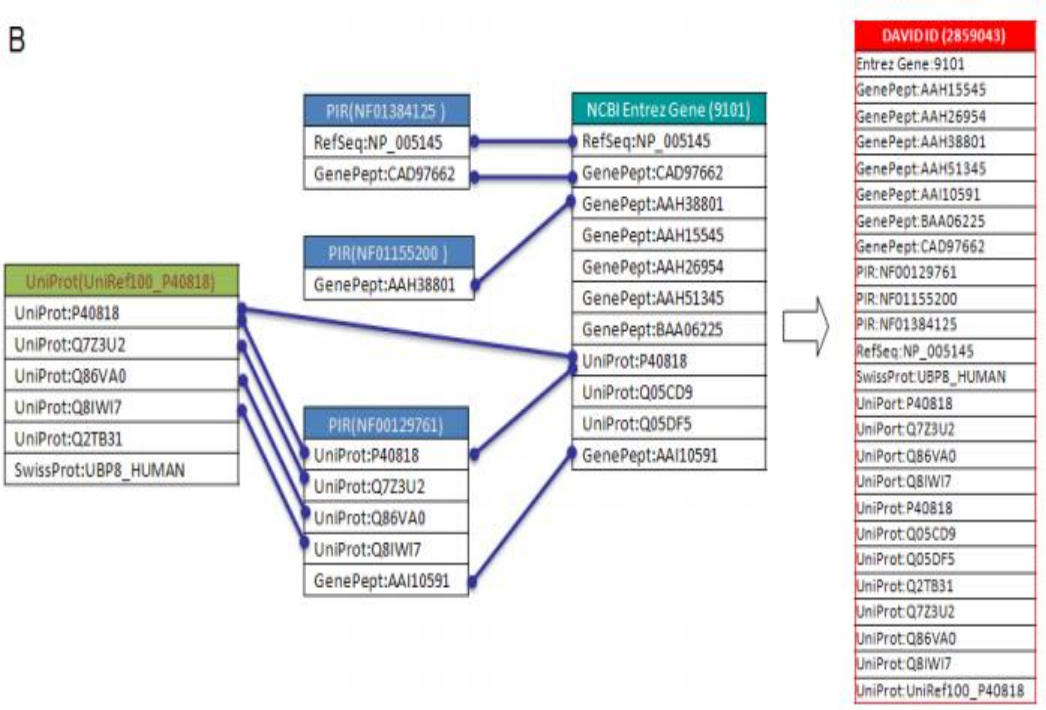


图 3.1　单联算法实现示例

通过单联算法本文最终筛选出有 484 条数据存在冗余现象，并将 484 条冗余基因 ID 凝聚为 219 条 DAVID ID。另外，原则上，无论是在正常情况下还是耐药情况下，一个基因的表达水平应该是要么上调，要么下调，而不存在说同一种情况下既有上调又有下调。因此，对于数据集中同一基因既有上调又有下调的数据本文亦选择放弃，所以，最终本文共筛掉 299 条冗余数据。

（2）标注错误

廖奇等人[46-47]经研究发现，exon array 中的一些探针被错误注释到 exon 上，即存在 LncRNA（长非编码 RNA）与 mRNA（信使 RNA）标记混乱的情况，因此为了防止本文所用初始数据集亦存在 LncRNA 与 mRNA 标记混乱的情况，本文将初始数据集与 NCBI 中 Gene 数据库中最新 Gene type 进行比对，最终筛选出 16 个存在标注错误的数据，而此类数据并不编码蛋白，因此本文亦将此类数据放弃。

经上述数据处理后得到的数据集为本文研究所用最终数据集，如何从这些数据中找到研究人员所关切的、有研究价值的数据。为此，本文首先对数据集进行分类，将其分

为若干类，然后再通过这些分类进行后续调控网络研究。

3.2　数据的注释分类

DAVID 提供了一套全面的基因注释工具，它集成了来自几十个公共数据库的 40 多个功能性分类的高质量注释内容[48]，它根据不同类型的基因 ID 从不同的注释数据库分配注释目录给相同的 DAVID 基因簇。DAVID 集中了两个成对的 TEXT 文件（基因索引文件，注释索引文件），然后通过它们来实现基因注释的分配。如图 3.2 所示，如查询 affy_id 207849_at (IL2)注释条目，首先通过基因索引文件获得相应的 DAVID 标识符（红色），然后通过该标识符在不同注释索引文件中顺序查询注释条目。

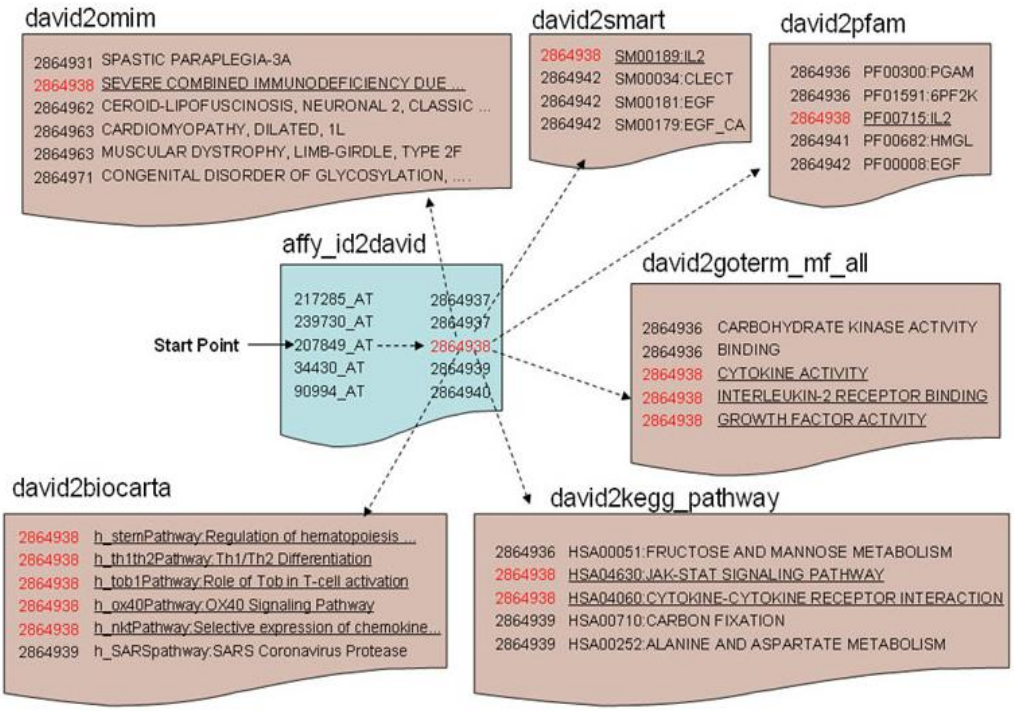


图 3.2　成对 TEXT 文件

借助于此方法，本文将数据集依据注释内容进行了初次分类，数据集共被分为了疾病、功能性分类、基因本体论、通路以及蛋白域五大类，而除疾病类外，其它四大类又分为三小类，具体参见图 3.3。其中，横轴为分类，左纵轴为每分类基因数占总基因数的百分比，右纵轴为每分类所拥有的基因数。

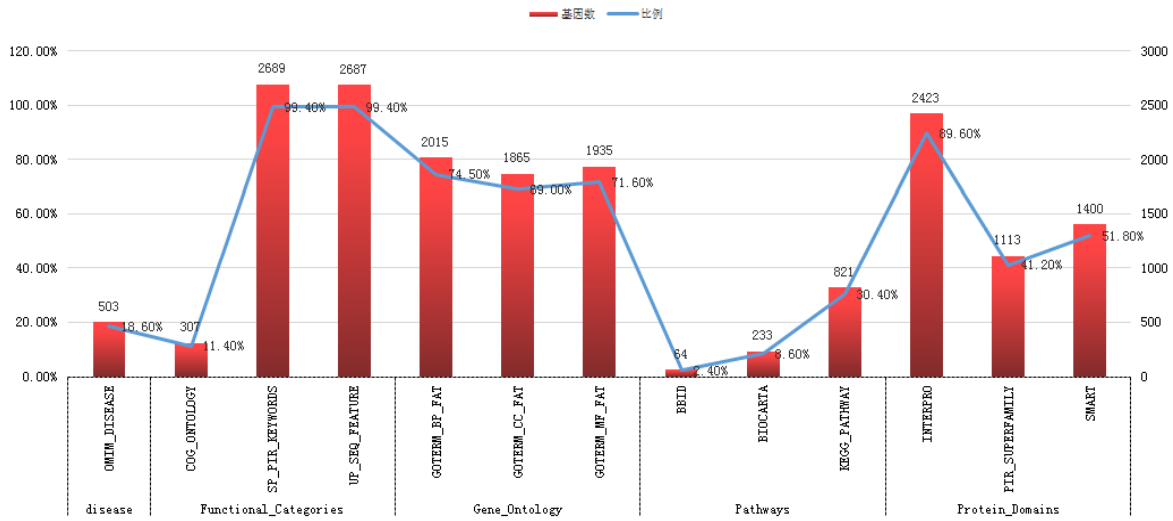


图 3.3　基因注释分类图

相关研究表明功能相关的基因总能聚集在一起，而相关基因的识别则可以通过基于两个基因有相似注释内容的假设衡量全局注释内容相似性实现。简言之，就是我们可以依据全局注释内容相似性来衡量基因对间的功能关系。越是相关的基因越会聚集在一起，因此为了进一步提高各分类数据间的相关性，本文在上述注释分类基础上对各分类数据又进行了一次功能性聚类，从而将每分类数据又分为若干个小规模类。

3.3　数据的功能性聚类

对于功能性聚类[49]，首先采用一个方法来衡量基因间的相似性，随后再使用聚类算法依据相似距离对相关基因进行聚集，算法中加入了模糊特征，该特征使得一个基因可以属于一个或多个聚集群，其过程可由以下流程体现：

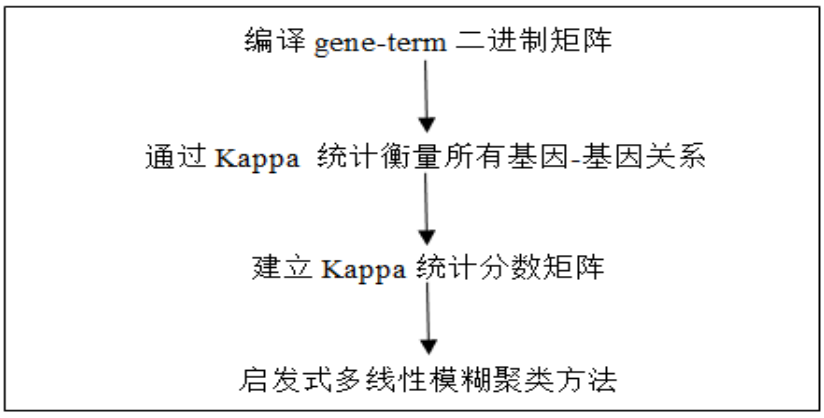


图 3.4　功能性聚类流程

由上述流程可知，第一步，编译 gene-term 二进制注释矩阵。设想将所有 terms 在一个平面线性集合上划分为独立的 term，每个基因都与一些注释 term 相关，从而构建 gene-term 二进制矩阵，如表 3.1 所示，1 表示该基因有当前 term，0 表示未知。

表 3.1　gene-term 二进制矩阵

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
Gene a	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Gene b	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Gene c	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
Gene d	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0
Gene e	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Gene f	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Gene g	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
Gene h	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0

第二步，依据上述建立的 gene-term 矩阵，采用 Kappa 统计衡量任何给定基因对间的注释同现。公式 3.1 则为 Kappa 值计算公式，其中 p，q 分别代表基因 p 和基因 q，O_{pq} 代表观察到的同现，A_{pq} 代表可能的同现，K_{pq} 为 Kappa 值，它代表基因 p，q 之间注释同现的度，值为 1 表示同现性好，为 0 表示同现与随机一样。

$$K_{pq}=\frac{O_{pq}-A_{pq}}{1-A_{pq}} \quad \text{……（3.1）}$$

现针对表 3.1 中的基因 a，d 给出 Kappa 值的计算过程：

		Gene a: 1	Gene a: 0	总行
Gene d	1	6（$C_{1,1}$）	1（$C_{0,1}$）	7（$C_{1,*}$）
	0	4（$C_{0,1}$）	4（$C_{0,0}$）	8（$C_{0,*}$）
	总列	10（$C_{*,1}$）	5（$C_{*,0}$）	15（T_{ad}）

$$O_{ad}=\frac{C_{1,1}+C_{0,0}}{T_{ad}}=\frac{6+4}{15}=\frac{2}{3}$$

$$A_{ad}=\frac{C_{*,1}\times C_{1,*}+C_{*,0}\times C_{0,*}}{T_{ad}\times T_{ad}}=\frac{10\times 7+5\times 8}{15\times 15}=\frac{22}{45}$$

$$K_{ad}=\frac{O_{ad}-A_{ad}}{1-A_{ad}}=\frac{\frac{2}{3}-\frac{22}{45}}{1-\frac{22}{45}}=\frac{8}{23}=0.35$$

基因两两之间经公式 3.1 计算后，便可得出第三步的 Kappa 统计分数矩阵，表 3.2

则是表 3.1 对应的 Kappa 统计分数矩阵。基因对间重叠注释越多，Kappa 分数越高，一般认为分数在 0.35 以上较好。

表 3.2　Kappa 统计分数矩阵

	a	b	c	d	e	f	g	h
a		1	1	0.35	-0.50	-0.50	-0.50	0.00
b	1		1	0.35	-0.50	-0.50	-0.50	0.00
c	1	1		0.35	-0.50	-0.50	-0.50	0.00
d	0.35	0.35	0.35		0.35	0.35	0.35	-0.11
e	-0.50	-0.50	-0.50	0.35		1	1	0.00
f	-0.50	-0.50	-0.50	0.35	1		1	0.00
g	-0.50	-0.50	-0.50	0.35	1	1		0.00
h	0.00	0.00	0.00	-0.11	0.00	0.00	0.00	

第四步，采用启发式多线性模糊聚类方法实现相关基因聚集，具体包括以下三步：1. 多个初始种子：每个基因只要它满足比三个以上基因的 Kappa 值大于 0.35 都可被选为初始聚簇中心，另外为了保证聚类质量，该聚簇中 50%以上成员彼此之间 Kappa 值也要满足大于等于 0.35。2. 最小多重线性法合并种子，当 2 个种子共享 50%以上成员时则进行合并。3. 返回步骤 2 继续合并，直到不能合并为止。

以下我们用图解法形象的描述该聚类方法（图 3.5）。（a）假想每个基因都位于一个虚拟的二维空间内，距离则代表了基因间关系（Kappa 值）的度。（b）初始种子群，即上述步骤 1。（c）迭代合并种子群直至没有合并发生，图中红框标注的三个基因为“孤儿基因”，它们不隶属于任何种子群。（d）最终形成三个基因簇，涂红基因则是因算法具有模糊特征，使得其可属于一个或多个聚集群。

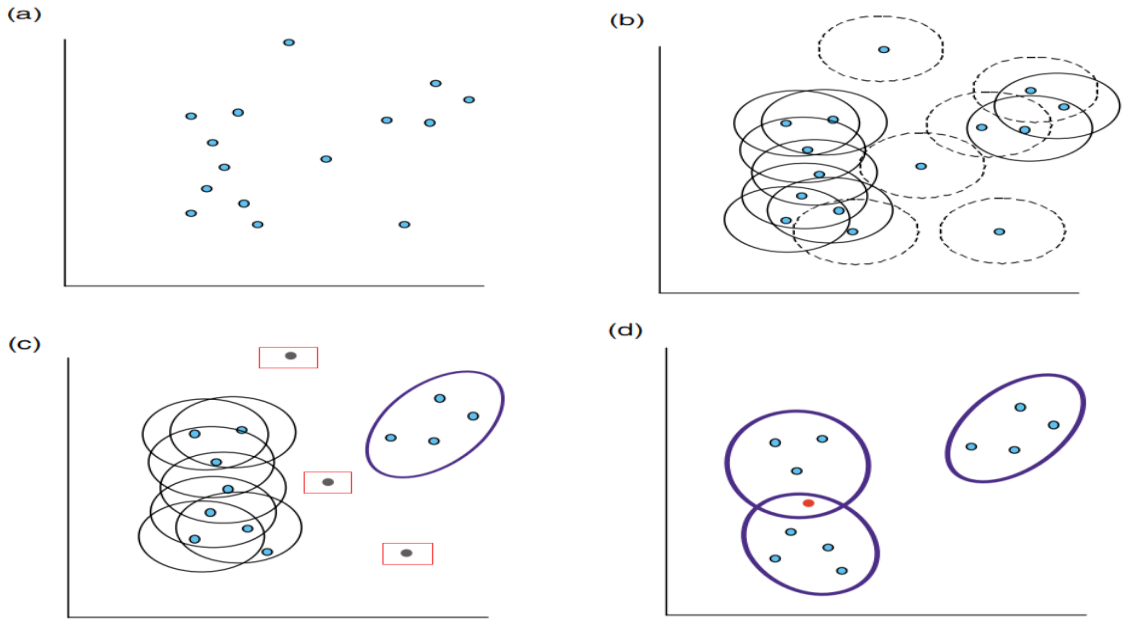


图 3.5　聚类方法的图解说明

各分类数据经功能性聚类后发现，除相关基因外还存在部分基因为“孤儿基因”或不相关基因，它们不映射到任何已聚集的基因群上，即图 3.5（c）红框所示，而导致这一现象产生的原因有三点：1. 它与任何其他基因都没有关系。2. 它与其他一些基因有关系，但因不满足最小聚集群成员数，使得其不能形成一个聚集群。3. 假阴性。因为此类数据与其他数据之间无相关性，所以此类数据亦放弃不予考虑。各注释分类数据排除“孤儿基因”或不相关基因后，便得出本文最终所用分类数据集。图 3.6 为初次分类与各分类删除不相关基因后基因数比对图。

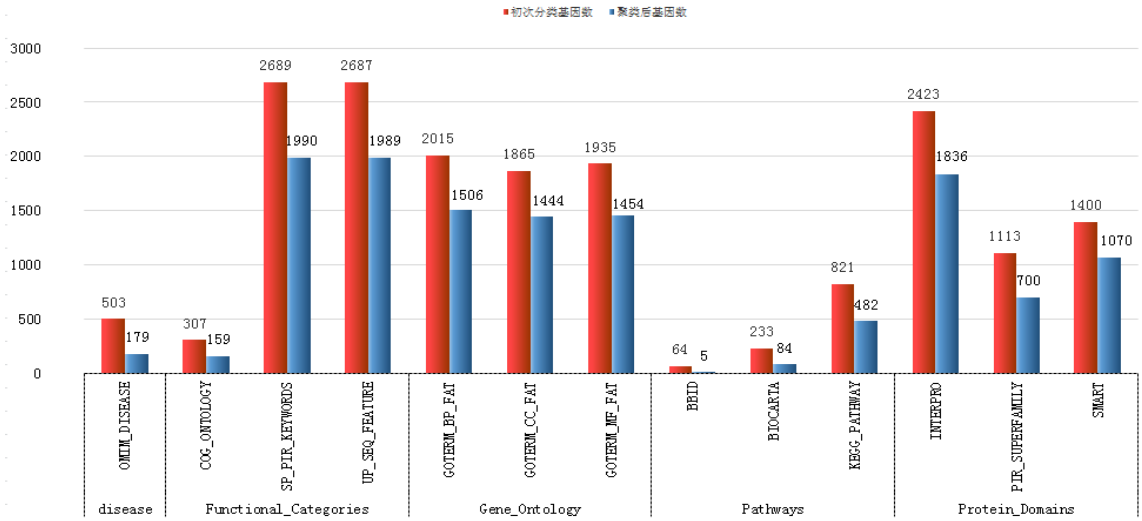


图 3.6　最终基因分类图

3.4　本章小结

本章主要介绍了数据集差异表达基因的预处理工作。首先，本文从数据冗余和标注错误两方面对数据集中存在的不确定错误或噪声进行了处理。随后，本文依据注释内容将处理后数据初次分类。最后，为了进一步提高分类后数据间的相关性，本文依据全局注释内容相似性衡量基因对间的功能关系理论，在初次分类基础上再进行功能性聚类，将每类数据再分为若干个小规模类，同时在聚类分析时发现，存在部分基因不映射到任何已聚集的基因群上的情况，对于此类基因为了保障数据集的质量本文亦选择放弃。

第 4 章　基于胃癌耐药数据的调控网络研究

4.1　调控网络总体流程的设计

原始的差异表达基因数据经上述预处理后，便得到了本章用于构建调控网络模型的最新数据。为此，本文先对预处理后的数据进行融合，建立初级调控网络，然后用衡量网络节点重要性的评分机制对初级调控网路中的各节点进行打分，并依据分值对节点排序，随后依据节点分值选取适当的节点作为核心节点，并以这些节点为中心设计网络挖掘模块方法，最后对挖掘出的核心网络模块进行分析，最终找到与胃癌 MDR 相关的有价值数据。其总体流程可概括为以下四步，总体流程可见图 4.1：

（1）预处理后差异表达基因集融合，建立初级调控网络。

（2）网络节点重要性打分，节点排序，以及选取适当核心节点。

（3）依据核心节点挖掘网络模块。

（4）对挖掘出的核心网络模块进行分析。

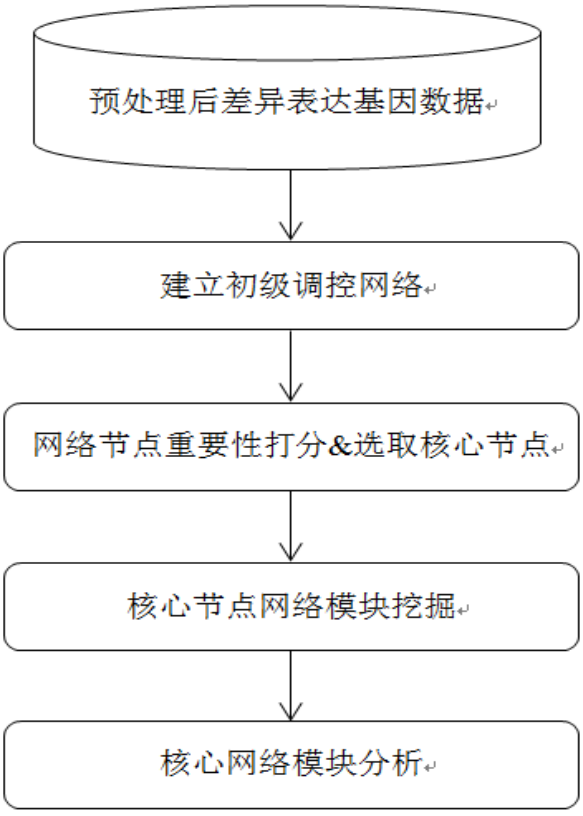


图 4.1　调控网络总体流程

4.2　调控网络的具体设计

4.2.1　建立初级调控网络

图形理论，简称图论，目前已被广泛用于细胞内部分子间相互作用结构特征的识别，而这些理论也可以很好地帮助我们定量地描述生物系统的网络结构。用最简单的图论术语来说，一个网络由一系列节点和节点间的连线组成，其中，节点可以是任何学者们所感兴趣的东西，如 DNA、RNA、基因、代谢物等，而节点间的连线则表示两个相连节点间存在作用关系。依据节点间作用关系的性质，可将图形分为有向图和无向图两种。有向图中，如图 4.2 所示，节点间的作用关系有明确定义方向，比如新陈代谢中物质的流动方向。无向图中，如图 4.3 所示，节点间的作用关系不具有方向性，如蛋白相互作用网络。

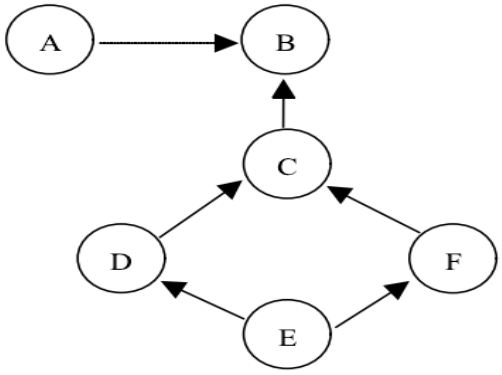


图 4.2　有向图

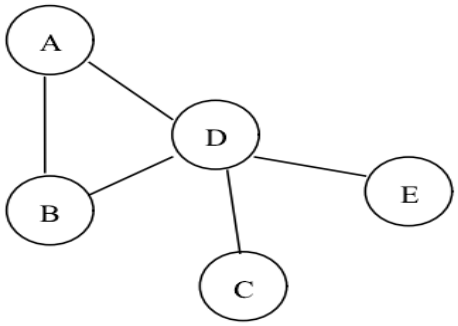


图 4.3　无向图

调控网络，作为对 DNA 片段、RNA 分子，以及各类蛋白间相互作用描述的一种生物网络，它以系统的观点为出发点，通过对网络中的各分子及其间的作用关系研究来探索生命现象。因此，可以用图论来描述调控网络，所以，本文将预处理后的差异表达基因数据进行融合，建立了初级调控网络，从而把各基因间的作用关系以网络图这种更为直观的形式展现，以进行进一步研究。

在本文里，网络中的节点代表某一个基因，因此，本文的调控网络也称为基因调控网络，节点的边代表两个基因间的作用关系。另外，因为基因与基因间的作用关系是相互的，因此本文选用无向图。基因与基因间的相互作用关系是建立网络的纽带，因此，找到基因间的相互作用关系，从而组织成初级基因调控网络是本小节的下一步工作。

经过之前阶段对差异表达基因数据的错误或噪声处理，注释分类，以及功能性聚类

后得到的最新数据，为调控网络的初级建立做了基础准备。此阶段，本文选用了生物信息学软件 GENEMANIA，用以挖掘预处理后的各分类差异表达基因间的相互作用关系，从而建立初级调控网络。

GENEMANIA[50-51]，一个提供用户基因或蛋白功能预测的分析网站，其数据库包括 1800 个网络，涵盖了 8 个器官的 50 万个相互作用关系，8 个器官分别为：拟南芥，线虫，果蝇，斑马鱼，人类，肌肉，鼠类，酵母菌，相互作用关系则分为直接作用和间接作用两种，其中直接作用包括 Co-expression（共表达）、Pathway（共享通路）、Physical interactions（物理相互作用）、Co-localization（共定位）、Genetic interactions（遗传相互作用）和 Shared protein domains（共享蛋白结构域）六种，而间接作用只有 Predicted（通过一个媒介）一种。

由 GENEMANIA 挖掘出的各分类基因间的相互作用关系经相关处理后，便可得出各分类相应的基因数，以及相互作用关系数，即节点数和边数，详见表 4.1。

表 4.1　各分类基因数、相互作用关系数情况

分类		基因节点数	相互作用边数
disease	OMIM_DISEASE	155	301
Functional_Categories	COG_ONTOLOGY	144	1529
	SP_PIR_KEYWORDS	1888	24889
	UP_SEQ_FEATURE	1887	25150
Gene_Ontology	GOTERM_BP_FAT	1424	17103
	GOTERM_CC_FAT	1373	18690
	GOTERM_MF_FAT	1386	17380
Pathways	BBID	4	6
	BIOCARTA	81	182
	KEGG_PATHWAY	454	2780
Protein_Domains	INTERPRO	1692	22515
	PIR_SUPERFAMILY	658	4264
	SMART	1034	14047

节点数和边数便组成了初级调控网络，图 4.4，图 4.5 是使用 Cytoscape 软件做出的分类数据集中 SP_PIR_KEYWORDS 和 UP_SEQ_FEATURE 的初级调控网络图，以图这

种直观的形式展示这两个初级网络。在这里，因为 SP_PIR_KEYWORDS 和 UP_SEQ_FEATURE 分类均涵盖了总基因数中 99%以上的基因，所以给出了这两个分类的初级网络图，其他分类初级网络图与它们相比只是节点数和边数不同，因此不再一一给出。

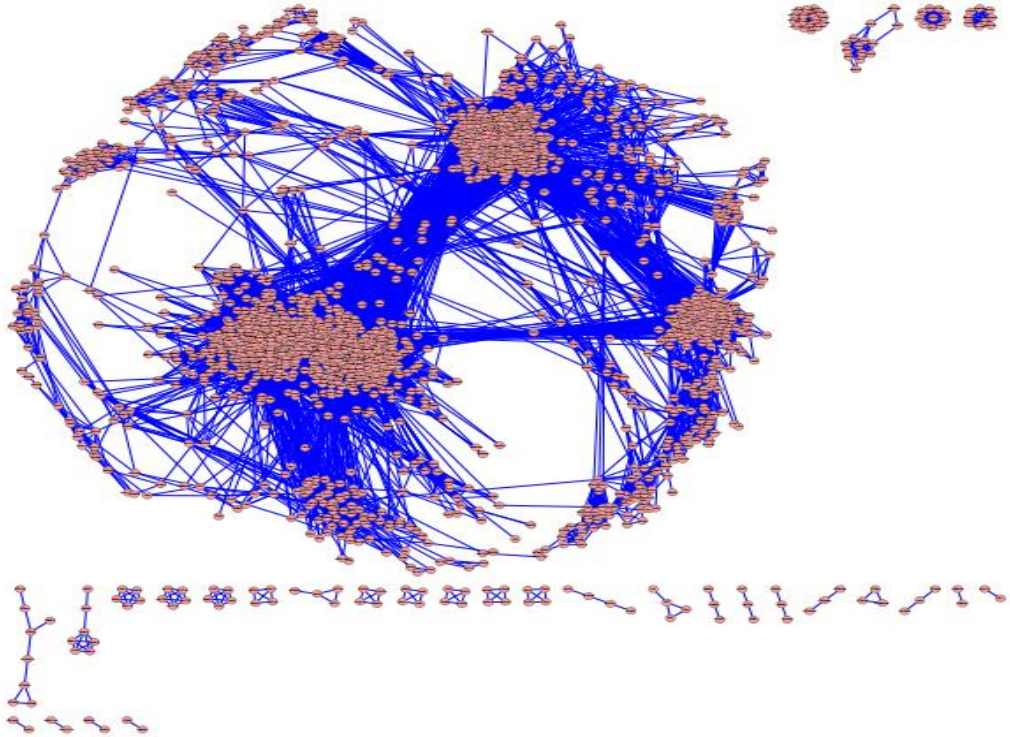


图 4.4　SP_PIR_KEYWORDS 初级调控网络图

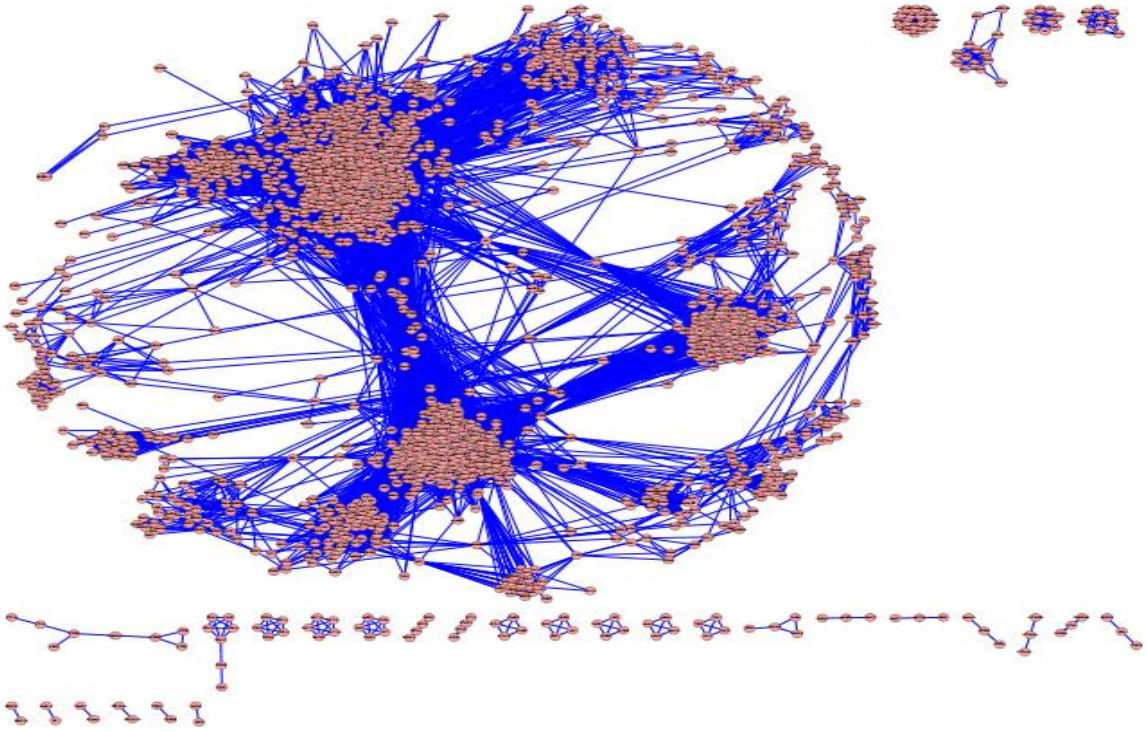


图 4.5　UP_SEQ_FEATURE 初级调控网络图

4.2.2　网络节点重要性打分，选取核心节点

在调控网络中，尽管节点数的变化范围广泛，但这些节点在网络中并不是起同等作用的，通常处于网络中心的节点与其重要性成正比关系，也就是说，越是处于网络中心的节点其重要性越高。因此，对初级基因调控网络中的节点进行打分排序以更好的分析调控网络，从而挖掘有价值的数据显得尤为重要。在这里，本文首先对初级基因调控网络的各个节点进行打分评估，然后依据各节点的分值高低对节点排序，进而依据分值排名选取适当的节点作为下一步核心研究节点。

本文选用的初级调控网络节点打分评估算法为 PageRank[52-54]。该算法是 Google 用来对万维网中的网页进行排名的一种算法，在万维网中一个页面相当于一个网络节点，一个页面到另一个页面的链接则相当于网络中的一条边，而像雅虎、新浪等门户网页常常是人们浏览其他网页时需要通过的页面，这说明这些门户网页处于网络中心地带，地位相比其他网页也比较重要。因此，可以发现初级调控网络与万维网有着相似的节点和链接理解方式，所以可以把 PageRank 算法应用到本文的初级基因调控网络分析中去。

在 PageRank 算法中，每个页面均由一个反映其重要性的值，即 PR 值。而通常，影响页面 PR 得分值的因素包括：

（1）指向该网页的页面数量。

（2）指向该网页的页面自身的 PR 值。

（3）指向该网页的页面自身的链出数量。

其中，因素（1）、因素（2）与 PR 分值成正比，而因素（3）与 PR 分值成反比。

在初始阶段，首先用各网页间的链接关系建立 Web 图，并为每个页面分配相同的 PR 值，然后进行二维矩阵迭代相乘，直到所有页面分值达到稳定状态为止，即系统收敛，从而得出各个网页的最终得分。将此算法应用到本文的初级调控网络分析中，各基因节点就相当于一个个独立的页面，而基因间的相互作用关系便相当于各网页间的链接关系，其 PR 值计算过程如下：

$$A = q * P + \left(1 - q\right) * ee^{t} / n \quad \text{（4.1）}$$

$$B = \left(b_1, b_2, ..., b_i, ..., b_n\right) \quad \text{（4.2）}$$

$$B_i = A * B_{i-1} \quad \text{（4.3）}$$

（1）依据基因间的相互作用关系得出邻接矩阵 M。若两个基因之间有连接，则矩

阵 *M* 相应位置值为 1，反之值为 0。

（2）依据邻接矩阵 *M* 计算其概率转移矩阵 *P*。*P* 是将 *M* 的每行的每个元素除以其所在行所有元素之和，然后再将其转置得出的。

（3）依据概率转移矩阵 *P*，阻尼系数 *q*，逃脱因子 *1-q*，矩阵 e^t、*e* 以及基因节点总数 *n* 计算矩阵 *A*。其中，*q* 表示当前基因节点以 0.85 的概率移动到其相邻节点上，而 *1-q* 表示当前基因节点以 0.15 的概率移动到一个随机基因节点上，*e* 为全 1 的 1*n 矩阵，e^t 则为 *e* 的转置矩阵。

（4）初始时为每个基因节点分配相同的 PR 值 1，即矩阵 *B*。

（5）用矩阵 *B* 与矩阵 *A* 迭代相乘计算各基因节点在各时刻的 PR 值，经过若干次迭代后，矩阵 *B* 便可收敛，从而得出初级调控网络中各基因节点的 PR 值。

各分类差异表达基因数据经上述步骤计算后，便可得出相应的 PR 分值，由于结果集较大无法将计算的结果一一列出，所以在此本文只列出了分类数据集 SP_PIR_KEYWORDS 的前 15 名的 PR 值计算结果作为结果样例进行展示。

表 4.2　SP_PIR_KEYWORDS PR 分值表

基因名	节点的度	节点的 PR 值
RPRM	162	4.4039607
FAT4	128	3.5615509
RNF150	131	3.3406591
PSG1	102	3.031504
PHOX2B	96	2.9684677
RNF182	113	2.932518
SLC4A3	105	2.8450422
SMNDC1	26	2.8178637
CPE	106	2.7768376
GPM6A	106	2.6999986
AGBL1	101	2.6178246
CNTFR	101	2.6059382
CADM1	100	2.5950255
ROBO1	101	2.5884278
RNF43	95	2.5014248

对调控网络各节点打分之后，发现存在这样一些节点：PR 分值很高，但度却很小，如表 4.2 中的基因 SMNDC1，PR 分值排名第八，但节点的度却为 26，且其在度排名中居 758 名，这说明了本文所用的 PageRank 算法是以与网络节点的关系程度为标准来衡量节点重要性的，而不是以节点的连接数多少来衡量的。同时还发现，按 PR 分值排序后的末几名或末几十名的节点的度均为 1，也就是说，这些节点的重要性不是很高，所以，综合上述发现，本文对各分类数据中 PR 分值排名靠后，且度为 1 的节点不再予以考虑，去除这些节点后剩余的按 PR 分值高低排序的节点则为本文选取的核心节点，并用这些节点进行下一步研究。

4.2.3　核心节点网络模块挖掘

研究表明，某一蛋白常常与其附属蛋白组成一个蛋白模块，从而参与细胞生理活动，行使特定生物功能。通过前一阶段对各分类初级调控网络中各基因节点的打分、排序和分析后，便可选出各分类的核心节点，然后本文以选取的这些核心节点为中心进行模块挖掘，从而找到这些核心节点以及与它们附属的其他节点所组成的核心网络模块，进而对挖掘出的核心网络模块进行分析，最终找到与胃癌 MDR 相关的有研究价值的数据。

由前可知，本文的初级基因调控网络是由基因（节点）和基因间的相互作用关系（边）组成的无向图，它可用 $G=<V, E>$ 表示，其中 V 为基因集，E 为基因间的相互作用关系集。核心模块的挖掘则是在图 G 中寻找包含核心基因节点 V_k 的子图，即 $G_k=<V_k, E_k>$，这些子图需满足下述三个条件：

（1）子图中与核心节点 V_k 相邻的节点数必须大于 1，即核心节点的度大于 1。

（2）子图中的其他节点 V_i 必须与核心节点 V_k 直接相邻，即 V_k 与 V_i 存在边 E_{ki}。

（3）子图中与核心节点 V_k 相邻的节点 V_j 必须与其他和 V_k 相邻的节点 V_i 相邻，即 V_k 与 V_j 存在边 E_{kj}，V_k 与 V_i 存在边 E_{ki}，那么 V_j 与 V_i 也必须存在边 E_{ji}。

图 4.6 为本文设计的核心网络模块挖掘方法的流程图：

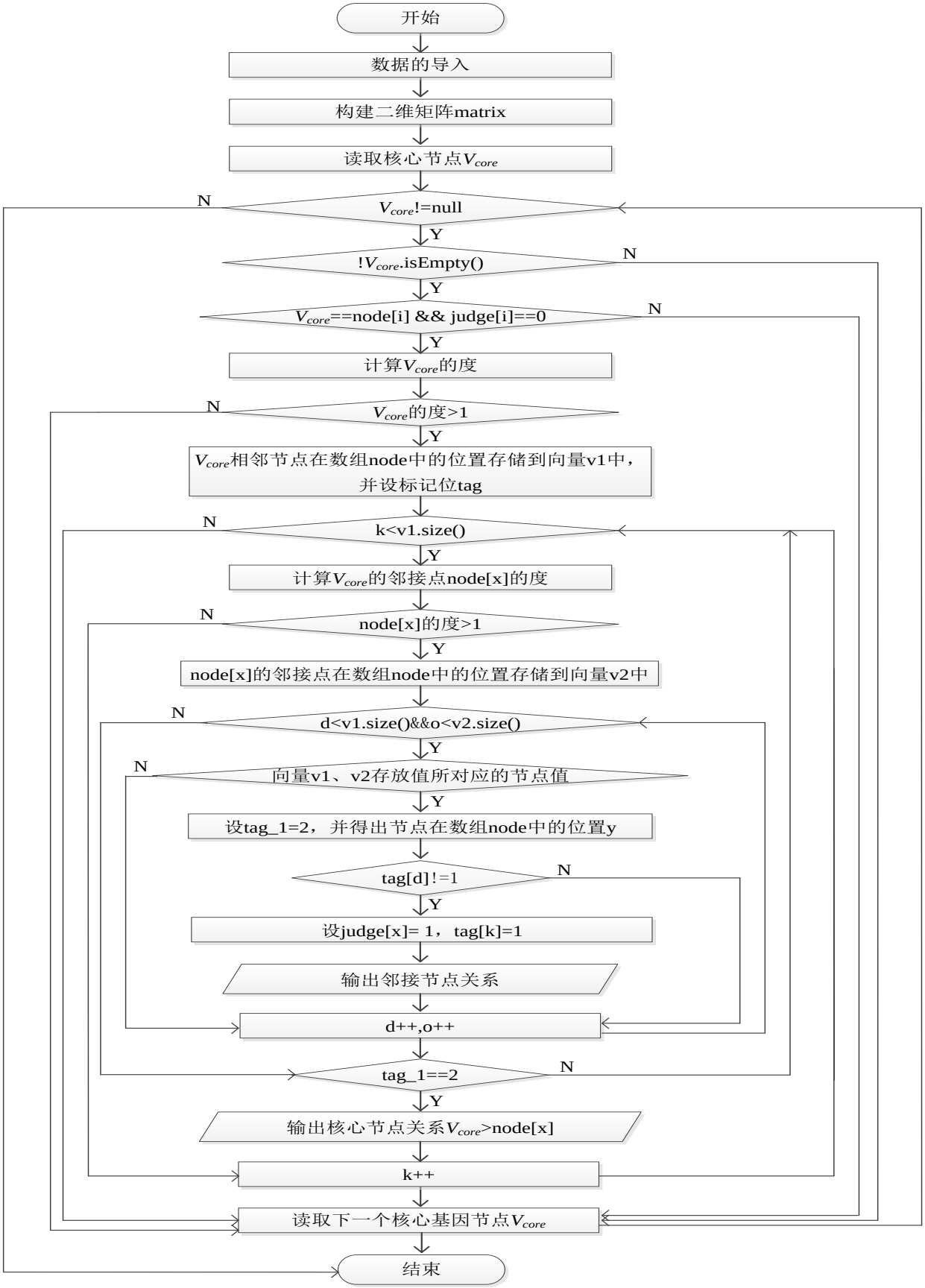


图 4.6　核心网络模块挖掘方法流程图

下面给出核心网络模块挖掘方法的具体实现步骤：

（1）输入数据的导入，本文的输入数据主要包含四个 txt 文件，分别为核心节点、初级调控网络节点、节点边关系的第一列和第二列。

（2）面向对象语言 java 建立 CoreModule 类，类中定义了一维数组 node、二维数组 matrix 分别用以存储节点信息和节点关系，另外，为了判断核心节点是否在其他核心模块中存在，本文定义了一维判断数组 judge，其初始值设为 0。

（3）依据节点间的关系构建二维矩阵 matrix，两个节点之间有连接值为 1，反之则为 0。

（4）读取核心节点 V_{core}，判断它是否与 node[i]（i=1,2,…,n）相等，并判断 judge[i]（i=1,2,⋯,n）是否为 0，即该核心节点是否在其他核心模块中已存在。若上述条件均满足，则计算 V_{core} 的度。当 V_{core} 的度大于 1 时，便将 V_{core} 的所有直接相邻节点在数组 node 中的位置存储到向量 v1 中，度小于 1 的节点不予考虑是为了降低节点间相关度低的现象。为了确保类似 A>B、B>A 的输出只有一个，定义了一维数组 tag 用来为向量 v1 的值设置标记，初始值设为 0。

（5）设置标记位 tag_1，初始值为 0，随后对 V_{core} 的每个邻接点 node[x]（x 是 V_{core} 的邻接点在数组 node 中的位置）做以下处理：

a）计算邻接点 node[x]的度，当 node[x]的度大于 1 时，便将 node[x]的直接邻接点在数组 node 中的位置存储到向量 v2 中；

b）判断向量 v1 和 v2 中存放的值所对应的节点是否相等，若条件为真，则得出节点在数组 node 中的位置 y，并设 tag_1 值为 2，表示需要输出；同时，当 tag[d]（d 为向量 v1 对应的位置）不为 1 时，就输出节点关系 node[x] > node[y]，并设 judge[x]为 1，表示节点 node[x]已存在于某核心模块中，同时设 tag[k]的值为 1，确保 A>B、B>A 的输出只有一个。

c）判断 tag_1 值是否为 2，若为真，表示邻接点模块输出完毕，需要输出核心节点 V_{core}>node[x]。

（6）读取其他核心节点返回步骤（4）继续处理，直至所有核心节点处理完毕。

各分类基因数据集选取的核心基因节点经过上述核心模块挖掘方法处理后，便可挖掘出各分类的核心节点的核心网络模块，其部分结果形式见图 4.7，其中，A 面板是挖掘出的核心节点模块，一个 txt 文件代表一个网络模块，文件名则以核心基因节点的名字命名，B 面板则是挖掘出的网络模块节点之间的作用关系。挖掘出核心节点网络模块

之后，需要找到哪些核心模块与胃癌 MDR 相关，具有研究价值，因此本文对挖掘出的核心网络模块又进行了相关分析。

图 4.7　核心网络模块挖掘结果形式

4.2.4　核心网络模块分析

对核心网络模块进行分析的主要目的是为了最终找到与胃癌 MDR 相关的有价值数据，为此，本文通过在 NCBI 中输入胃癌、胃癌耐药以及多药耐药等关键词统计了近 10 年已经通过生物实验验证的，与胃癌、胃癌耐药和多药耐药相关的基因，并将统计的这些基因与本文挖掘出的核心模块的各基因节点进行比对，最终发现有 16 个已经有研究证明与胃癌、胃癌耐药以及多药耐药相关的基因与本文挖掘出的网络模块中的节点一致，而这 16 个基因恰为网络模块的核心节点，随后本文对这 16 个核心基因节点以及其附属基因节点的注释说明进行了分析总结，发现这 16 个核心模块中以 ABCB1 为核心的网络模块是一个多药耐药模块，即该网络模块中的所有节点均与肿瘤的多药耐药相关，而其他核心网络模块中的部分节点也已经有研究证明它们与人类的前列腺癌、乳腺癌、头颈癌、口腔癌、结肠直肠癌、非小细胞肺癌，以及脑癌等多种疾病相关。

此外，在分析的过程中还发现有 4 个核心基因节点的分值排名与上述 16 个核心基因节点相近，而且目前也已有文献证明它们与肿瘤相关，因此，推断以这 4 个核心基因节点为中心的网络模块在胃癌的 MDR 中也起到了一定作用。所以，本文得出这 20 个核心基因节点与其附属基因节点所组成的网络模块在胃癌的 MDR 中起关键性作用。

表 4.3 是 20 个核心基因节点的分值及与其相关的文献报道的情况总结。这 20 个核心基因节点的分值排名相对较高，进一步说明了它们的重要性。其中，绿色底色部分为推断的 4 个核心基因，无底色部分为已证明的与胃癌及胃癌耐药相关的 16 个核心基因。

表 4.3　核心基因节点分值及文献报道总结

核心节点	PR 分值	与癌症相关的文献
RPRM	4.4039607	Endocrinology, 2012, 153(7): 2963-2973
FAT4	3.5615509	Cancer biology & therapy, 2015 (just-accepted): 00-00.
NTSR1	2.7708824	Chinese materia medica, 2015, 40(13): 2524-2536
TSPAN1	2.3427582	Tumori, 2008, 94(4): 531
SCN2A	2.2383697	Die Pharmazie-An International Journal of Pharmaceutical Sciences, 2015, 70(6): 416-420
GDF15	2.0835288	Future Oncology, 2014, 10(7): 1187-1202
FLT4	1.81831	Oncotarget, 2015, 6(5): 3225
TRIM24	1.8057698	Virchows Archiv, 2015, 466(5): 525-532
RUNX1	1.7835392	Biochemical and biophysical research communications, 2014, 448(3): 315-322
NCOA1	1.7552673	Oncotarget, 2015, 6(27): 23890-23904
CXCR4	1.7288183	Anticancer research, 2014, 34(11): 6397-6403
SEMA3A	1.5549979	International journal of clinical and experimental pathology, 2013, 7(8): 4782-4794
GABRA1	1.3041713	Die Pharmazie-An International Journal of Pharmaceutical Sciences, 2015, 70(6): 416-420
ARNT	1.2802396	PloS one, 2014, 9(6): e99242
ACSS2	1.070627	Journal of surgical oncology, 2015, 112(6): 585-591
AREG	1.0580546	Pharmacogenetics and genomics, 2014, 24(11): 539-547
ERBB3	1.0115926	Acta biochimica et biophysica Sinica, 2014, 46(3): 190-198
ANXA2/ANXA3	1.0000186/ 1.0000044	Biomedicine & Pharmacotherapy, 2015, 69: 237-241/ Neoplasma, 2013, 61(3): 257-264
ABCB1	1.0000097	New England Journal of Medicine, 2003, 348(15): 1442-1448
CD82	1.0000005	Digestive diseases and sciences, 2015, 60(7): 1967-1976

以下是对多药耐药核心网络模块 ABCB1 各基因的注释说明：

ABCB1，它是最早被发现且研究也最为深入的 ABC 转运蛋白超家族成员之一，细胞无论是对药物的摄取还是排出都会涉及到细胞膜上的运输或者转运通道，而 ABC 转运蛋白便具有这样一类通道。ABC 分为 ABC1、MDR/TAP、CFTR/MRP、ALD、OABP、GCN20、White 7 个亚科。ABCB1 作为 ABC 亚家族 B（MDR/TAP）成员 1，其主要作用是减少细胞内药物的累积，以使得肿瘤细胞因此获得耐药性。ABCB1 主要参与了药物响应、跨膜转运蛋白两个生物过程，是细胞成分、膜组分、质膜等细胞组件的重要成分，有核苷酸结合、蛋白结合、水解酶活性等多种分子功能。

TAP1，又称 ABCB2，是参与多药耐药亚家族 B（MDR/TAP）成员 2。ABCB2 参与了蛋白的转运、跨膜转运、蛋白定位等生物过程，是内质网、细胞质膜、细胞部分等细胞组件的成分，有核苷酸结合、ATP 结合、水解酶活性、ATP 酶活性等多种分子功能。

ABCB4，亚家族 B（MDR/TAP）成员 4，而 MDR/TAP 亚科的成员主要参与多药耐药性及抗原递呈，且该基因可能也涉及从肝脏肝细胞到胆汁磷脂的运输。ABCB4 参与了内源性刺激物的响应、器官的响应等生物过程，是细胞膜、质膜组件的组成部分，有核苷酸结合、ATP 结合、多药转运活性、主要活性跨膜转运等多种分子功能。

ABCC2，参与多药抗性亚家族 C（CFTR/MRP）成员 2，其主要在肝细胞的小管部分和胆道运输中表达，底物主要包括长春新碱等抗癌药物，所以，该基因通常有助于药物的耐药性。ABCC2 参与了分子的运输生物过程，是质膜、膜组件的组成部分，有核苷酸结合、ATP 结合、转运活性、ATP 酶活性以及跨膜物质运动等多种分子功能。

ABCC4，参与多药抗性的亚家族 C（CFTR/MRP）成员 4，在细胞排毒中居重要地位。ABCC4 参与了离子转运和跨膜转运两个生物过程，是细胞质膜、囊泡膜以及细胞质囊膜等细胞组件的组成部分，有核苷酸结合、ATP 结合、水解酶活性等多种分子功能。

ABCG1，亚家族 G（White）成员 1，它参与了巨噬细胞胆固醇和磷脂的转运，而且可能在其他细胞类型中调控脂质的平衡。ABCG1 参与了糖蛋白分解代谢过程、脂质转运、糖蛋白转运、转录调控等生物过程，是核内体、内质网、高尔基体、质膜、细胞表面等细胞组件的组成部分，有核苷酸结合、ATP 结合、水解酶活性、蛋白二聚化活性等多种分子功能。

ABCD3，亚家族 D（ALD）成员 3。ABCD3 参与了脂质运输、有机酸转运、跨膜转运等生物过程，是细胞质体膜、细胞器膜等细胞组件的成分，有核苷酸结合、ATP 结合、水解酶活性、ATP 酶活性等多种分子功能。

由上可以看出，核心网络模块 ABCB1 各节点均与多药耐药相关，16 个核心节点又

都与胃癌、胃癌耐药以及多药耐药相关，而推测的 4 个核心节点也都与肿瘤相关。所以综合上述分析可以得出在 20 个起关键作用的核心网络模块中，20 个核心基因和 ABCB1 多药耐药网络模块又在其中发挥了主要作用。

4.3　研究结果

挖掘出的核心网络模块经分析后，本文得出以 ABCB1、RUNX1、CXCR4、ERBB3、GABRA1、ANXA2/ANXA3、ACSS2、CD82、FAT4、AREG、ARNT、GDF15、SCN2A、SEMA3A、TRIM24、TSPAN1、RPRM、NTSR1、FLT4、NCOA1 这 20 个核心基因节点为中心的网络模块对胃癌的 MDR 起了关键性作用，具有一定的研究价值。而 20 个核心基因节点以及以 ABCB1 为核心基因的多药耐药网络模块又在其中发挥了主要作用。

表 4.4 是本文得出的 20 个核心网络模块的部分模块展示：

表 4.4　部分核心网络模块节点表

核心基因	网络模块所包含的其他基因
ABCB1	ABCC4、ABCB4、TAP1、ABCG1、ABCC2、ABCD3
RUNX1	THRB、TEAD1、ZFHX3、PPARG、MECOM
CXCR4	SGCE、ADRA2C、OLR1、CDH15、MME、CXCR1
ERBB3	FLT4、MYLK、PDK2、PRKAA1、ZAP70、WEE1、INSRR、PRKCB、SIK2、TESK1、ABL2、CDK6、MAP3K15
GABRA1	CHRNA7、CHRNB1、CHRNB4
ANXA2/ANXA3	ANXA6、ANXA13、ANXA3/ANXA2
ACSS2	PKM、MVD、KIF5B、CAMK2D、PDK2、MYO16、KIF13A、SIK2、ETNK2、STK38、ACSM3、NEK6、FGFR2、PC、KALRN、ACSL3、SLC27A2、ACSL6、ACSL5、TESK1、THNSL1、SMARCAD1、MYH14、ACSL1、ALPK2
CD82	TSPAN16、CD151、TSPAN4
NCOA1	FOXO1、MITF
FLT4	ICK、FGFR2、FGFR3、CDKL5、ACVRL1、LMTK2、CDK6、CD79B、ERBB3、ROR2、IL12RB1

在挖掘出的网络模块中发现，两个以 ANXA2 和 ANXA3 为中心的网络模块，它们的节点数和边数完全一致，唯一区别是一个以基因 ANXA2 为中心，一个以基因 ANXA3 为中心，所以本文将这两个模块合并为了一个，即 ANXA2/ANXA3，并将这两个节点均作为模块的核心节点。

图 4.8–4.17 是表 4.4 中所列出的核心网络模块的图形化展示，网络模块图则由 Cytoscape 软件做出。模块图中的各基因节点分别用红色和蓝色表示，红色节点表示基因在胃癌耐药细胞中相比正常细胞而言，其表达水平为上调，而蓝色则表示表达水平为下调，另外，为了突显核心基因与其附属基因的不同，各模块图中核心基因的尺寸相对其他基因较大。

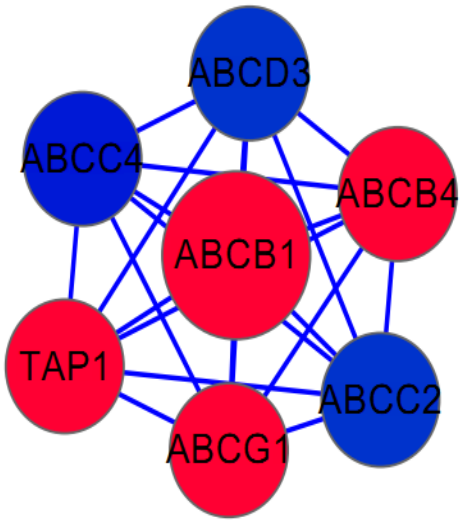


图 4.8　ABCB1 网络模块

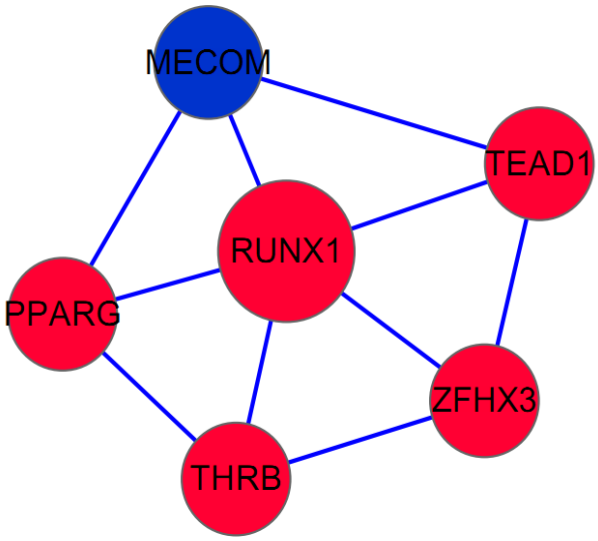


图 4.9　RUNX1 网络模块

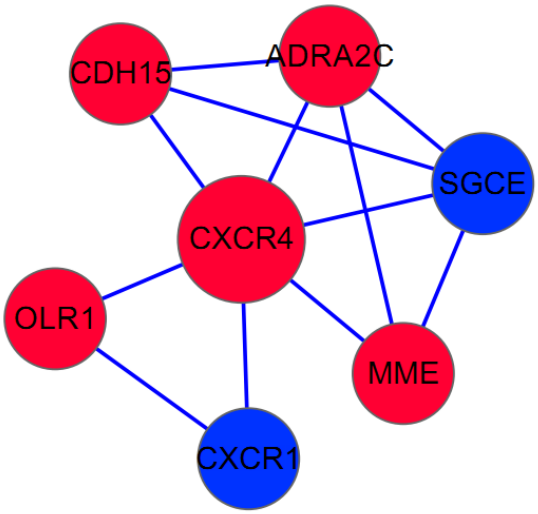


图 4.10　CXCR4 网络模块

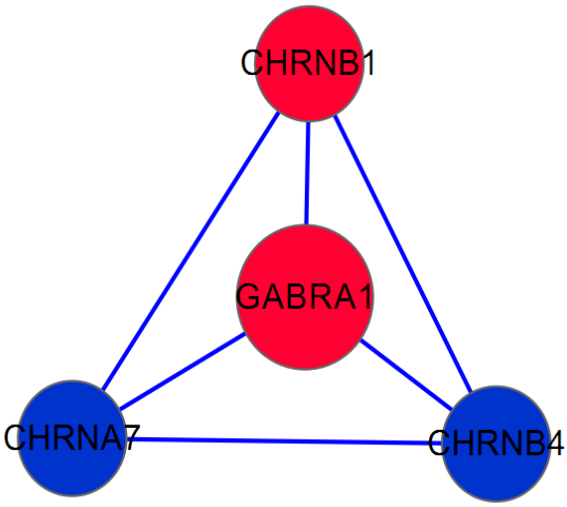


图 4.11　GABRA1 网络模块

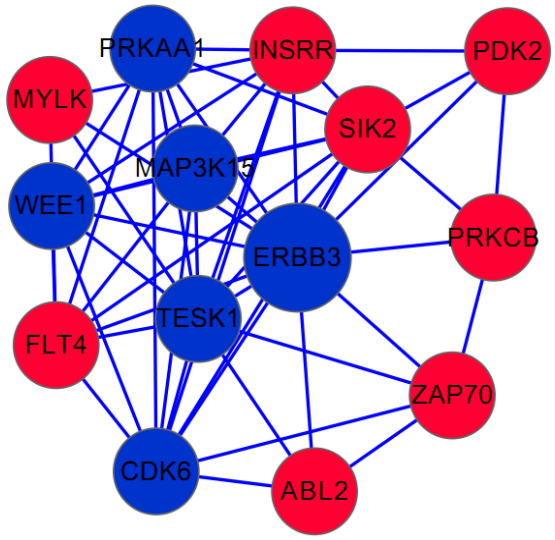


图 4.12　ERBB3 网络模块

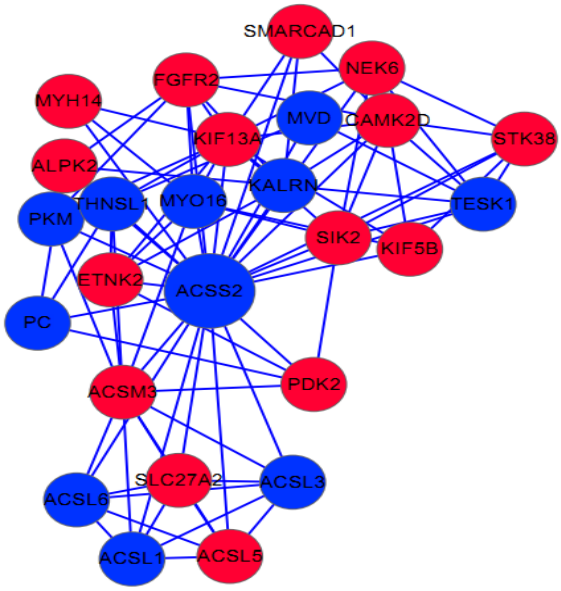


图 4.13　ACSS2 网络模块

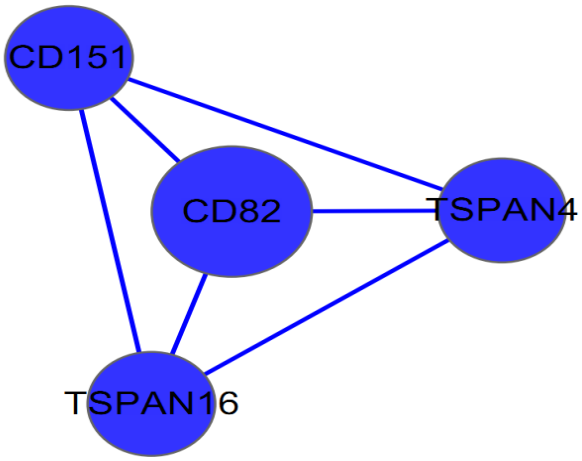


图 4.14　CD82 网络模块

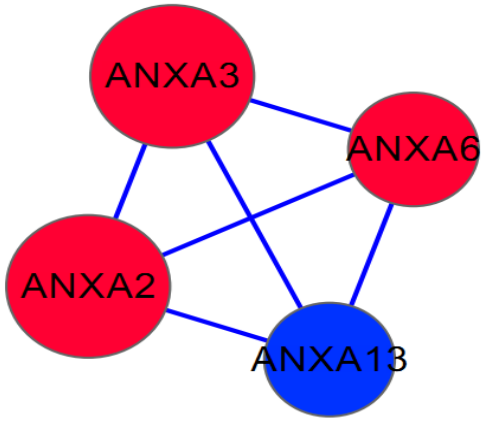


图 4.15　ANXA2/ANXA3 网络模块

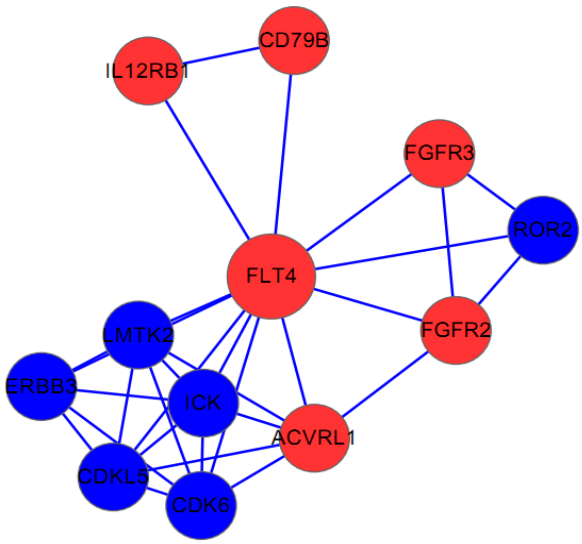


图 4.16　FLT4 网络模块

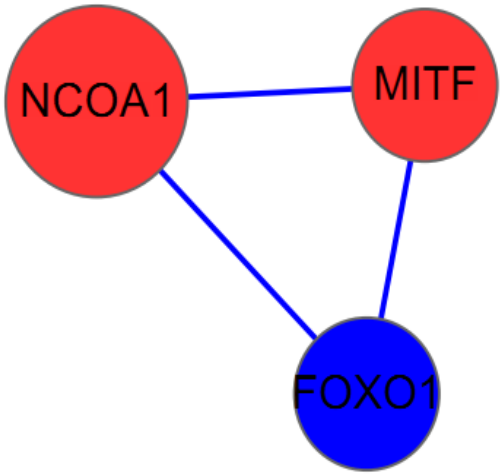


图 4.17　NCOA1 网络模块

4.4　模型有效性评估

为了进一步评估本文提出的基于胃癌耐药的调控网络生物信息学模型的有效性，本文用另一套数据对该模型方法进行了测试评估，这套数据来源于 HPRD（人类蛋白参考数据库），该数据库主要用于研究蛋白质交互关系，其数据一般认为是真实的数据。该套蛋白数据经融合处理后建立的 PPI（蛋白交互）初级网络如图 4.18 所示。

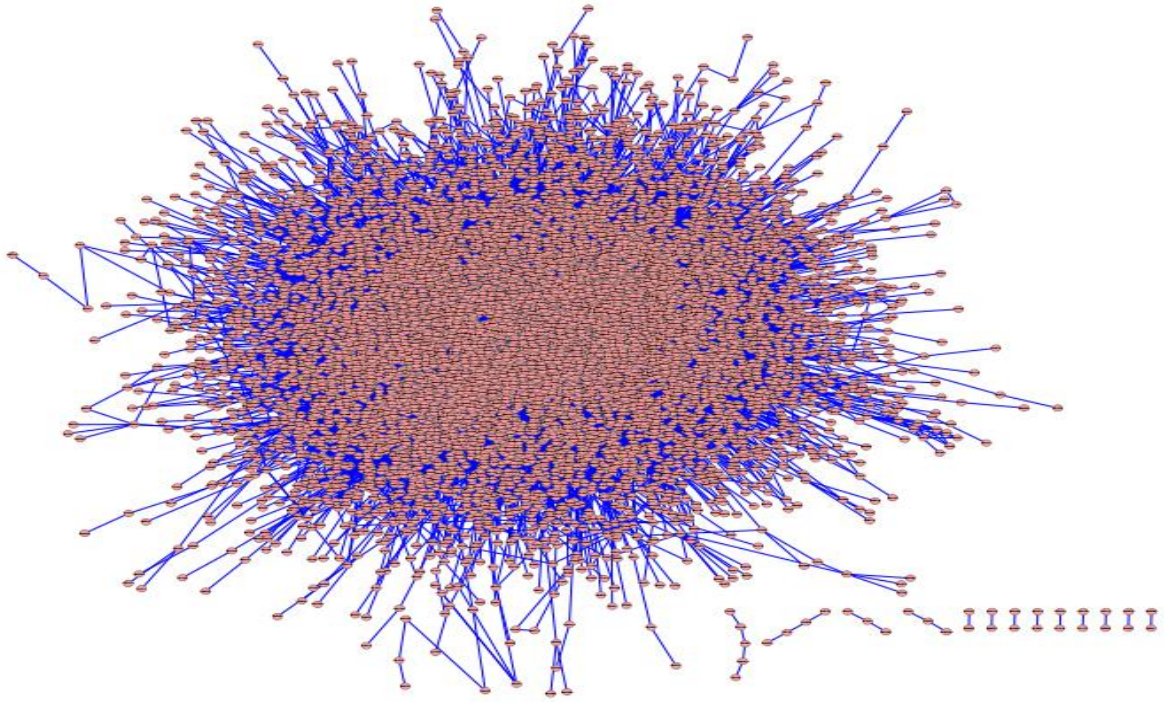


图 4.18　PPI 初级网络

黑色素瘤，一类常见的皮肤黏膜、色素膜恶性肿瘤，是皮肤肿瘤中恶性程度最高的瘤种，而目前在 UniProt（全球蛋白质资源）中记载的与黑色素瘤相关的蛋白有 200 多种，且这些蛋白全部包含在 PPI 初级网络中。

该套蛋白数据用本文提出的模型方法处理、分析后，能有效挖掘出以 P55072、P26599、P04075、P06733、P43364、P25024 这 6 个核心节点为中心的并与黑色素瘤相关的蛋白网络模块，而这 6 个蛋白网络模块的核心蛋白以及其部分附属蛋白都高度与黑色素瘤相关。因此，该套蛋白数据的分析结果有效证明了本文提出的基于胃癌耐药的调控网络生物信息学模型的有效性，也从侧面证明了本文所得结论的有效性。

图 4.19–4.24 为挖掘出的 6 个与黑色素瘤相关的蛋白网络模块。

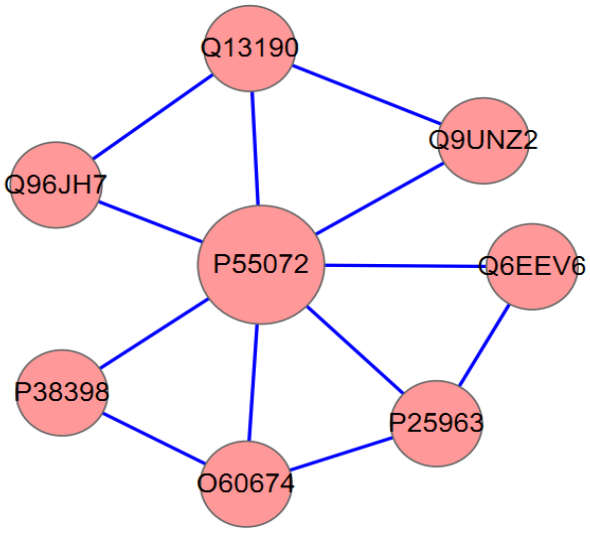


图 4.19　P55072 网络模块

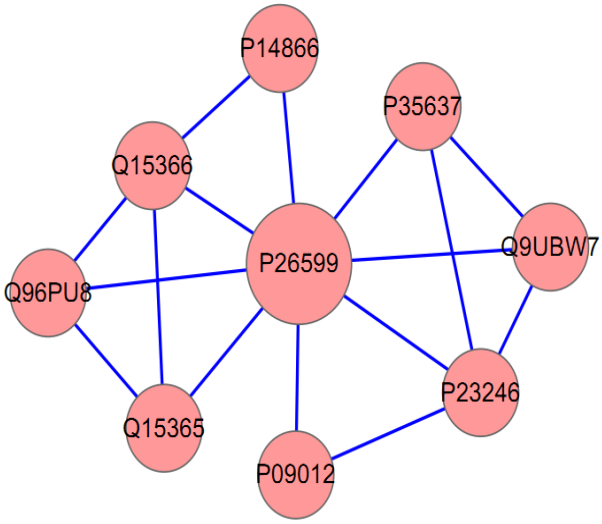


图 4.20　P26599 网络模块

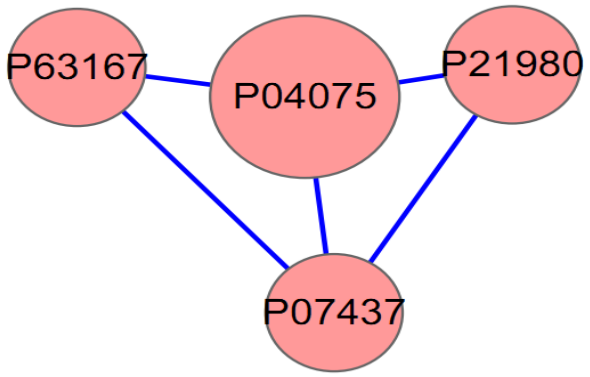


图 4.21　P04075 网络模块

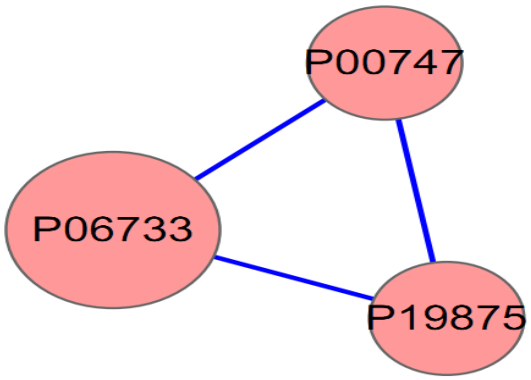


图 4.22　P06733 网络模块

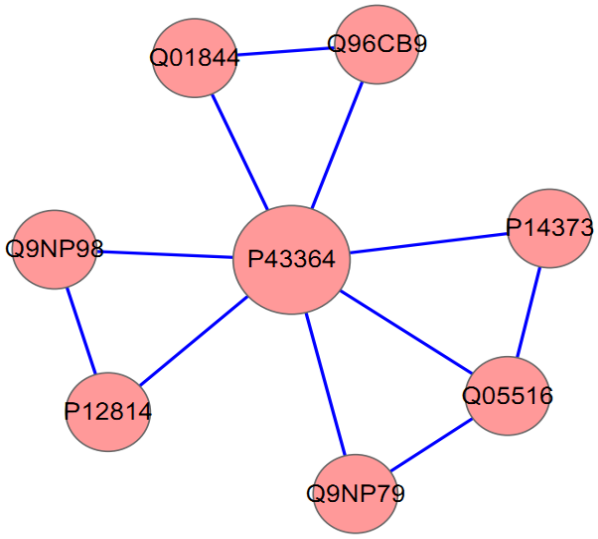


图 4.23　P43364 网络模块

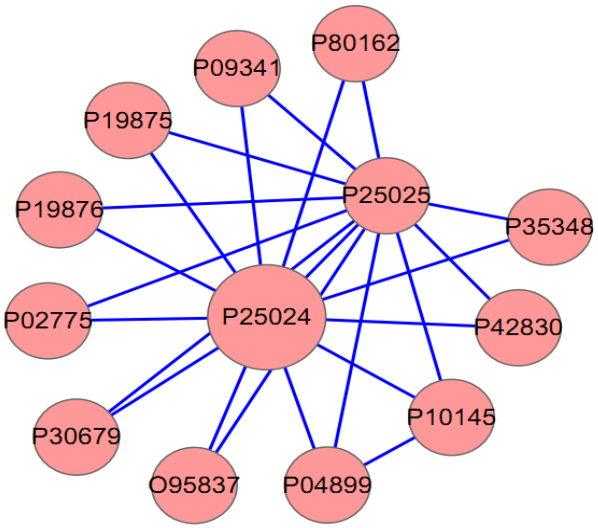


图 4.24　P25024 网络模块

4.5　本章小结

本章主要对基于胃癌耐药数据的调控网络研究做了详细的介绍。首先，对调控网络研究的总体流程进行了设计，然后按照所设计的流程一步步进行详细的设计研究，从而获得最终的研究结果。

流程包括以下四步：第一步，对第三章预处理后的各基因数据进行融合，从而建立初级调控网络，该阶段本文借助了生物信息学软件工具 GENEMANIA；第二步，网络中各基因节点的重要性不是同等的，所以本文对第一步所构建的初级网络中的各节点进行了打分，打分方法则使用了 PageRank 算法，随后依据打出的分值对节点排序，然后再依据节点分值选取适当的核心节点进行下一步研究；第三步，节点通常通过与其附属节点组成一类模块，从而行使相关生物功能，所以该阶段本文以第二步所选取的核心节点为中心挖掘这些核心节点以及它们的附属节点所组成的网络模块；第四步，分析挖掘出的核心网络模块。

经相关分析后，本文得出 20 个核心基因以及它们的附属基因所组成的网络模块对胃癌的 MDR 起了关键性作用，具有一定的研究价值，而 ABCB1 多药耐药网络模块和 20 个核心基因又在其中起了主要作用。

最后，为了进一步评估本文提出的模型方法的有效性，本文用另一套蛋白数据对其进行了测试评估，分析结果有效证明了本文提出的生物信息学模型方法的有效性，同时也从侧面证明了本文所得结论的有效性。

第 5 章　总结与展望

5.1　总结

目前针对胃癌 MDR 的研究多以生物学实验为主，而生物学因受人力、物力、成本等因素的限制使得研究的基因数量往往有限，而这对于涉及多基因，多因素以及多步骤的胃癌 MDR 研究而言显然远远不够。为此，本文借助生物信息学方法，通过合理的建模并结合网络模型，对胃癌耐药差异表达基因数据进行相关统计分析、筛选、挖掘，以从中找到与胃癌 MDR 相关的有研究价值的数据。

（1）对合作单位提供的胃癌耐药差异表达基因数据集进行预处理。首先，本文从数据冗余和标注错误两类问题上对初始数据集中存在的不确定错误或噪声进行了处理，随后依据注释内容对处理后数据进行初次分类，最后，为了进一步提高分类后数据间的相关性，本文在初次分类基础上用聚类算法对数据集再进行功能性聚类，从而将每类数据再分为若干个小规模类。

（2）在前期功能性聚类后数据基础上，对数据进行融合，然后通过本文挖掘出的基因间的相互作用关系建立各分类初级基因调控网络，随后又设计了评分机制，对各分类初级基因调控网络中的各节点进行打分、排序，最后依据各节点分值选取适当的核心节点。

（3）以选取的这些核心节点为中心，设计核心基因调控网络模块挖掘方法。接着，对挖掘出的以核心节点为中心的核心模块进行相关分析，经分析后，本文得出 20 个核心网络模块与胃癌的 MDR 相关，具有研究价值。其中，以 ABCB1 为中心的多药耐药模块和 20 个核心基因节点发挥了主要作用。

这种借助生物信息学方法，从网络模型角度出发，基于胃癌耐药的调控网络计算模型研究，既为胃癌 MDR 生物实验提供了理论指导，也为寻求胃癌耐药机理提供了新的思维方式。同时，本文也对促进胃癌耐药机理研究、临床治疗、转录组学研究等科学领域都有着重要的理论与现实意义，对其他癌症耐药研究也具有重要借鉴作用。

5.2　展望

本文提出的基于胃癌耐药的调控网络计算模型研究，是首次结合网络模型，针对胃癌 MDR 的生物信息学研究模型。本文的研究数据主要为差异表达基因，最终通过核心基因调控网络模块挖掘方法，找到与胃癌 MDR 相关的有价值数据，然而，人类整个基因组中还存在诸如 siRNA（小干扰 RNA）、miRNA（小分子 RNA）和 LncRNA（长非编码 RNA）等这样一类并不编码蛋白的 RNA，ncRNA（非编码 RNA）。而这些非编码 RNA 往往在转录水平上会调控基因的表达，进而影响基因的表达量，如何将这些非编码 RNA 引入到本文的胃癌耐药调控网络中，从而使得建立的调控网络更具准确性则是本文下一步需要改进的地方。另外，本文的网络模块挖掘方法也存在诸多不足亟需进一步改进以更好的挖掘有价值数据。最后，本文所得出的结果仍需要最终通过生物实验验证其有效性。但本文相信，该基于胃癌耐药的调控网络生物信息学研究模型，以及本文的研究结果将对实验人员的胃癌 MDR 研究发挥一定的参考借鉴作用。

参考文献

[1] Wang R, Chen X Z. High mortality from hepatic, gastric and esophageal cancers in mainland China: 40 years of experience and development[J]. Clinics and research in hepatology and gastroenterology, 2014, 38(6): 751-756.

[2] Liang J, Ge F, Guo C, et al. Inhibition of PI3K/Akt partially leads to the inhibition of PrPC-induced drug resistance in gastric cancer cells[J]. FEBS journal, 2009, 276(3): 685-694.

[3] Zhao L, Pan Y, Gang Y, et al. Identification of GAS1 as an epirubicin resistance-related gene in human gastric cancer cells with a partially randomized small interfering RNA library[J]. Journal of Biological Chemistry, 2009, 284(39): 26273-26285.

[4] Hendrickson DG, Hogan DJ, Herschlag D, et al. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance[J]. PloS one, 2008, 3(5): e2126-e2126.

[5] Alberts SR, Cervantes A, Van de Velde CJH. Gastric cancer: epidemiology, pathology and treatment[J]. Annals of Oncology, 2003, 14(90002): 31-36.

[6] Fan K, Fan D, Cheng LF, et al. Expression of multidrug resistance-related markers in gastric cancer[J]. Anticancer research, 1999, 20(6C): 4809-4814.

[7] Zhao Y, You H, Liu F, et al. Differentially expressed gene profiles between multidrug resistant gastric adenocarcinoma cells and their parental cells[J]. Cancer letters, 2002, 185(2): 211-218.

[8] Wang X, Lan M, Shi YQ, et al. Differential display of vincristine-resistance-related genes in gastric cancer SGC7901 cell[J]. WORLD JOURNAL OF GASTROENTEROLOGY, 2002, 8(1): 54-59.

[9] 赵屹, 谷瑞升, 杜生明. 生物信息学研究现状及发展趋势[J]. 医学信息学杂志, 2012, 33(5): 2-6.

[10] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of molecular biology, 1970, 48(3): 443-453.

[11] Pipas JM, McMAHON JE. Method for predicting RNA secondary structure[J]. Proceedings of the National Academy of Sciences, 1975, 72(6): 2017-2021.

[12] Lareau C A, White B C, Oberg A L, et al. Differential co-expression network centrality

and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure[J]. BioData mining, 2015, 8(1): 1.

[13] Wang L, Feng Z, Wang X, et al. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data[J]. Bioinformatics, 2010, 26(1): 136-138.

[14] Zheng H, Hang X, Zhu J, et al. REMAS: a new regression model to identify alternative splicing events from exon array data[J]. BMC bioinformatics, 2009, 10(Suppl 1): S18.

[15] Suphavilai C, Zhu L, Chen J Y. A method for developing regulatory gene set networks to characterize complex biological systems[J]. BMC genomics, 2015, 16(Suppl 11): S4.

[16] Kanehisa M. Post-genome informatics[M]. Oxford University Press (OUP), 2000.

[17] Shmulevich I, Saarinen A, Yli-Harja O, et al. Inference of genetic regulatory networks via best-fit extensions[M].Computational and Statistical Approaches to Genomics. Springer US, 2002: 197-210.

[18] D'haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering[J]. Bioinformatics, 2000, 16(8): 707-726.

[19] Bansal M, Belcastro V, et al. How to infer gene networks from expression profiles[J]. Molecular systems biology, 2007, 3(1): 78.

[20] Chuang CL, Chen CM, Shieh GS, et al. A fuzzy logic approach to infer transcriptional regulatory network in saccharomyces cerevisiae using promoter site prediction and gene expression pattern recognition[C]. Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on. IEEE, 2008: 1714-1721.

[21] Perina A, Seoane D, Gonzáez-Tizón A M, et al. Molecular organization and phylogenetic analysis of 5S rDNA in crustaceans of the genus Pollicipes reveal birth-and-death evolution and strong purifying selection[J]. BMC evolutionary biology, 2011, 11(1): 304.

[22] Simon-Loriere E, Rossolillo P, Negroni M. RNA structures, genomic organization and selection of recombinant HIV[J]. RNA biology, 2011, 8(2): 280-286.

[23] Chen T, He HL, Church GM. Modeling gene expression with differential equations[C]. Pacific symposium on biocomputing. 1999, 4(29): 4.

[24] Tominaga D, Koga N, Okamoto M. Efficient Numerical Optimization Algorithm Based on Genetic Algorithm for Inverse Problem[J].

[25] Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets[J]. Journal of theoretical biology, 1969, 22(3): 437-467.

[26] Shmulevich I, Dougherty E R, Kim S, et al. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks[J]. BIOINFORMATICS, 2002, 18(2):

261-274.

[27]Shmulevich I, Dougherty E R, Zhang W. From Boolean to probabilistic Boolean networks as models of genetic regulatory networks[J]. Proceedings of the IEEE, 2002, 90(11): 1778-1792.

[28]张晗, 宋满根, 陈国强, 等. 一种改进的多元回归估计基因调控网络的方法[J]. 上海交通大学学报, 2005, 39(2): 270-274.

[29]En Chai L, Saberi Mohamad M, Deris S, et al. Current Development and Review of Dynamic Bayesian Network-Based Methods for Inferring Gene Regulatory Networks from Gene Expression Data[J]. Current Bioinformatics, 2014, 9(5): 531-539.

[30]Beck W T. The cell biology of multiple drug resistance[J]. Biochemical pharmacology, 1987, 36(18): 2879-2887.

[31]Crick F. Central dogma of molecular biology[J]. Nature, 1970, 227(5258): 561-563.

[32]杨歧生. 分子生物学[M]. 浙江大学出版社, 2004.

[33]陈启民, 耿运琪. 分子生物学[M]. 南开大学出版社, 2001.

[34]De Jong H. Modeling and simulation of genetic regulatory systems: a literature review[J]. Journal of computational biology, 2002, 9(1): 67-103.

[35]张明菊. 基因表达的调控机制[J]. 黄冈职业技术学院学报, 2000, 1: 019.

[36]刘长宁, 孙世伟. 复杂生物网络及非编码 RNA 参与的双色网络[J]. 信息技术快报, 2010, 8(1):33-60.

[37]Karlebach G, Shamir R. Modelling and analysis of gene regulatory networks[J]. Nature Reviews Molecular Cell Biology, 2008, 9(10): 770-780.

[38]Lähdesmäki H, Shmulevich I, Yli-Harja O. On learning gene regulatory networks under the Boolean network model[J]. Machine learning, 2003, 52(1-2): 147-167.

[39]Politano G, Savino A, Benso A, et al. Using Boolean networks to model post-transcriptional regulation in gene regulatory networks[J]. Journal of Computational Science, 2014, 5(3): 332-344.

[40]Kim S Y, Imoto S, Miyano S. Inferring gene networks from time series microarray data using dynamic Bayesian networks[J]. Briefings in bioinformatics, 2003, 4(3): 228.

[41]Covert M W, Schilling C H, Palsson B. Regulation of gene expression in flux balance models of metabolism[J]. Journal of theoretical biology, 2001, 213(1): 73-88.

[42]McAdams H H, Arkin A. It's a noisy business! Genetic regulation at the nanomolar scale[J]. Trends in genetics, 1999, 15(2): 65-69.

[43]Gillespie D T. A general method for numerically simulating the stochastic time evolution

of coupled chemical reactions[J]. Journal of computational physics, 1976, 22(4): 403-434.

[44] Huang D W, Sherman B T, Tan Q, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists[J]. Nucleic acids research, 2007, 35(suppl 2): W169-W175.

[45] Huang D W, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nature protocols, 2008, 4(1): 44-57.

[46] Liao Q, Liu C, Yuan X, et al. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network[J]. Nucleic acids research, 2011, 39(9): 3864-3878.

[47] Liao Q, Xiao H, Bu D, et al. ncFANs: a web server for functional annotation of long non-coding RNAs[J]. Nucleic acids research, 2011, 39(suppl 2): W118-W124.

[48] Sherman B T, Huang D W, Tan Q, et al. DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis[J]. Bmc Bioinformatics, 2007, 8(1): 426.

[49] Alvord G, Roayaei J, Stephens R, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists[J]. Genome Biol, 2007, 8(9): 183.

[50] Montojo J, Zuberi K, Rodriguez H, et al. GeneMANIA: Fast gene network construction and function prediction for Cytoscape[J]. F1000 Research, 2014, 3(3):153-153.

[51] Warde-Farley D, Donaldson S L, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function[J]. Nucleic acids research, 2010, 38(suppl 2): W214-W220.

[52] Brin S, Page L. The anatomy of a large-scale hypertextual Web search engine[J]. Computer Networks & Isdn Systems, 1998, 30(98):107-117.

[53] Voevodski K, Teng S H, Xia Y. Spectral affinity in protein networks[J]. BMC systems biology, 2009, 3(1): 1.

[54] Grolmusz V. A note on the pagerank of undirected graphs[J]. Information Processing Letters, 2015, 115(6): 633-634.

作者简介

作者简介：

魏立艳，女，1987 年 03 月 25 日，汉族，河北省石家庄人，2013-2016 年就读于吉林大学软件学院，硕士研究生，研究方向为生物信息学。

获奖情况：

2013-2014 学年获得吉林大学研究生优秀二等奖学金以及优秀研究生荣誉称号。

致　谢

首先要感谢我的导师，刘元宁教授。因为他使我有幸进入生物信息学这个领域，并且也开始从事该领域的相关研究工作。感谢他在论文的撰写过程中对我的悉心指导，以及在论文构思和表达措辞上提出的一些宝贵意见。同时，在刘老师的指导之下，我也得以有幸参加国家自然科学基金、吉林省自然科学基金等项目的实践，并通过这些项目的参与，锻炼了我的沟通能力、组织能力，以及一定的科研能力。

在此同时，我还要感谢指导我日常研究的李志老师，感谢他对我工作的细心指导，他总会在我研究过程中提出一些新的想法，并在我遇到研究困难时，提出一些好的解决方法。更要感谢李志老师从课题开始研究到论文书写完毕各阶段所提出的宝贵意见。还要感谢张浩老师、段云娜老师、赵奇师兄他们对我学业上的帮助，因为他们使我更好的了解了该领域的研究。

同时，也要感谢我的亲人，感谢你们对我无微不至的照顾，以及一路的支持，你们的支持一直是我一步一步前进的动力。最后，感谢吉林大学软件学院为我们提供的良好学习环境，一流的师资队伍，并常年邀请国内外知名专家和学者做学术报告，大大的开阔了我的眼界。